DOCUMENT RESUME

ED 024 039 24 AL 001 598

By- Darnell, Donald K.

The Development of an English Language Proficiency Test of Foreign Students, Using a Clozentropy Procedure. Final Report.

Colorado Univ., Boulder. Dept. of Speech and Drama.

Spons Agency- Office of Education (DHEW), Washington, D.C. Bureau of Research.

Bureau No- BR-7-H-010

Pub Date Oct 68

Grant- OEG-8-8-070010-2000-057

Note- 73p.

EDRS Price MF-$0.50 HC-$3.75

Descriptors- *Cloze Procedure, College Students, *English (Second Language), *Foreign Students, Information Theory, *Language Tests, *Statistical Analysis, Testing, Test Interpretation, Test Reliability, Test Validity

Identifiers- *Clozentropy, Test Of English As A Foreign Language, TOEFL

This final report presents a description of a test combining cloze procedure and an entropy analysis (CLOZENTROPY), designed to measure the compatibility of a foreign student's English with that of his peers who are native speakers of English. This test, and the Test of English as a Foreign Language (TOEFL) were administered to 48 foreign students at the University of Colorado. (The CLOZENTROPY test was also administered to 200 native speakers of English at the same university.) Comparable reliability coefficients of approximately .86 were obtained for the two tests. Correlation between total scores on the two tests was .833. Analysis of variance confirms that content and difficulty of test material, major of subjects, and level and major of native comparison groups have significant influences on the CLOZENTROPY index of English proficiency. A discussion of the advantages over convention al types of tests and the major weakness (dependency on computer assistance in scoring), a sample copy of the test instrument, sample letters to the students, samples of computer output on the scoring program, and other data are included in the report. (AMM)

ED0 24039

BR 7-~~H~~-010
PA-24

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

FINAL REPORT

Project No. 7-H-010

Grant No. OEG 8-8-070010-2000 (057)

THE DEVELOPMENT OF AN ENGLISH LANGUAGE PROFICIENCY TEST OF FOREIGN STUDENTS, USING A CLOZENTROPY PROCEDURE

October 1968

U.S. DEPARTMENT OF
HEALTH, EDUCATION, AND WELFARE

Office of Education
Bureau of Research

AL 001 598

FINAL REPORT

Project No. 7-H-010

Grant No. OEG 8-8-070010-2000 (057)

THE DEVELOPMENT OF AN ENGLISH LANGUAGE PROFICIENCY TEST OF FOREIGN STUDENTS, USING A CLOZENTROPY PROCEDURE

Donald K. Darnell

University of Colorado

Boulder, Colorado 80302

October 1968

The research reported herein was performed pursuant to a grant with the Office of Education, U.S. Department of Health, Education, and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy.

U.S. DEPARTMENT OF
HEALTH, EDUCATION, AND WELFARE

Office of Education
Bureau of Research

TABLE OF CONTENTS

Preface. iv

Summary. 1

Chapter I Background. 2

A mathematical definition of abnormality-compatibility. Cloze procedure. Pilot studies.

Chapter II Sampling and Testing 12

The test passages. The criterion groups. Foreign student Ss. Test administration.

Chapter III Results 17

Reliability. Validity. Effects of six independent variables on CLOZENTROPY scores. ANOVA of native data. ANOVA of foreign data. Summary.

Chapter IV Summary and Conclusions. 28

References 34

Appendices

A. A sample copy of the test instrument 36
B. Sample letters to native students. 46
C. Sample letters to foreign students 51
D. Samples of computer output on scoring program. 55
E. Significantly different cell means--native data. 62
F. Significantly different cell means--foreign data 64

LIST OF TABLES

Table 1. Correlations among CLOZENTROPY, TOEFL, and GPA 18

Table 2. Analysis of variance of native student data. 20

Table 3. Analysis of variance of foreign student data 22

LIST OF FIGURES

Fig. 1 Level by content interaction (native) 20
Fig. 2 Major by content interaction (native) 21
Fig. 3 Major by level by content interaction (native). 21
Fig. 4 Content by difficulty by level interaction (native) 21
Fig. 5 Content by major interaction (foreign). 23
Fig. 6 Content by difficulty interaction (foreign) 23
Fig. 7 Content by rank of criterion group interaction. 23
Fig. 8 Major by rank of criterion group interaction. 24
Fig. 9 Content by major of criterion group interaction 24
Fig. 10 Difficulty by major of criterion group interaction. 24
Fig. 11 Content by major by rank of criterion group interaction . .24
Fig. 12 Difficulty by content by rank of criterion group interaction 25
Fig. 13 Difficulty by content by major of criterion group interaction 25
Fig. 14 Content by rank of criterion group by major of criterion group interaction 25
Fig. 15 Difficulty by rank of criterion group by major of criterion group interaction 26
Fig. 16 Major by content by difficulty by rank of criterion group interaction 26
Fig. 17 Content by difficulty by rank of criterion group by major of criterion group interaction. 26

PREFACE

This research report may be taken from two different points of view. Obviously, this is a validation study of a test instrument, but it may also be viewed as a validation study of a procedure for constructing test instruments. The procedure is an innovation in testing. It makes certain assumptions about language and its function and establishes an objective criterion related to those assumptions for the evaluation of individual linguistic performance.

The specific test created in this study is only an example of the kind of test that the CLOZENTROPY procedure produces. Although a copy of the test instrument appears in Appendix A, a scoring key does not appear in this report for three very good reasons: (1) The simplest hand scoring key imaginable would require the addition of approximately 150 pages to this report. A complete scoring key would require at least 300 additional pages. (2) Complete or incomplete, such a scoring key would have no practical value, because neither man nor machine could use it effectively. (3) The use of such a key by anyone not at the University of Colorado would be inconsistent with the theory underlying CLOZENTROPY procedure.

All the essential information which the individual user would need to develop his own CLOZENTROPY test and his own scoring key is provided in detail in this report. Hopefully, a computer program which would facilitate such developments will soon be published.

One could never acknowledge all the people who have contributed to a project of this kind, but there are six people who must be singled out for their contributions to this study.

First, my thanks go to Dr. David R. Saunders of the University of Colorado Department of Psychology for writing a computer program that made this project feasible. His encouragement and concrete assistance made possible the pilot research which led to the proposal for this study. His continued assistance in modifying the program for this specific application went far beyond what one can expect of a colleague.

My appreciation also goes to Mr. Glenn Bracht and Dr. Kenneth Hopkins of the University of Colorado Laboratory of Educational Research. Their advice on research design and assistance with the analysis contributed immeasurably to the success of this study.

The other three people whom I wish to thank are my chief assistant on the project, Stephen Clarke, and two other graduate students, Roger Babich and John Boyd, who volunteered their services. These three put in many hours of hard work and asked innumerable questions. Their work made the project possible, and their questions kept it interesting.

Of course, I take full responsibility for any errors that may have been made in the conduct of this research or in the writing of this report.

SUMMARY

The Test of English as a Foreign Language (TOEFL) and a CLOZENTROPY test were administered to a sample of 48 University of Colorado foreign students. The CLOZENTROPY test was also administered to 200 native speakers of English, students at the University of Colorado.

CLOZENTROPY procedure is a combination of cloze procedure and an entropy analysis derived from information theory. Its product is a measure of the compatibility of a foreign student's language patterns with the usage patterns among his native English speaking peers.

Comparable reliability coefficients of approximately .86 were obtained for the total scores on the two tests. Subtests in the two batteries were treated as items, and reliability coefficients were not obtained on the individual subtests.

Correlation analysis of the two tests indicates that the total scores on the two batteries are correlated .833. The communality accounts for almost all of the reliable variance in both of the tests. This is interpreted as positive support for the validity of the CLOZENTROPY test. Neither test was found to have any significant correlation with grade point average for the 48 foreign students tested.

ANOVA treatment of the data from 200 native students indicated significant differences between graduate and undergraduate students and between engineers and non-engineers. The level and major variables were found to interact with content and difficulty of the test material, such that the main effects are largely attributable to difficult material and engineering content.

ANOVA treatment of the data from 40 foreign students indicates no significant difference between graduate and undergraduate foreign students and no simple effects of the difference between engineering and non-engineering students. There were significant main effects of the message content variable and the two criterion group variables (level and major of native students). Interaction effects involve five of the six variables studied.

CLOZENTROPY procedure yields a reliable and valid test of English language proficiency that has a number of advantages over more traditional test procedures. Its one severe limitation is the need for computer assistance in scoring.

CLOZENTROPY procedure seems to be capable of producing highly specialized language tests for a variety of achievement levels and fields of interest. Further research along these lines is recommended.

THE DEVELOPMENT OF AN ENGLISH LANGUAGE PROFICIENCY TEST OF FOREIGN STUDENTS, USING A CLOZENTROPY PROCEDURE

Chapter I

Background

This is a report of a study conducted to test the reliability, validity, and practicality of a new test of English language proficiency. The new test employs a variation of Cloze Procedure (used previously to measure readability, comprehension, and language aptitude) and an entropy measure derived from information theory which indexes the compatibility of an individual's responses with those of a selected criterion group. The term, CLOZENTROPY, was coined to name this combination of procedures. "CLOZENTROPY test" or "CLOZENTROPY battery" will be used to refer to the particular test developed in this study, and "CLOZENTROPY procedure" will be used to refer to the general procedure which might be used to generate any number of specific tests.

The rationale or justification for this study is based on the following assumptions: (1) The primary function of language is communication. (2) This function is best served within any group by compliance with group norms of language usage. (3) A measure of proficiency in language should index the ability to conform to existing group norms of language rather than to some prescriptive model or idealized language pattern. (4) If language norms vary from group to group, the best measure of proficiency for an individual would be given in terms of the group or groups with whom he needs to communicate. (5) An ideal measure of language proficiency would take into account an individual's ability to exercise freedom of choice as well as his ability to comply with relatively rigid restrictions. That is, any natural language, in order to allow for new statements within the language structure, must permit users of that language some freedom of choice. Proficiency in the use of language, then, involves exercising that freedom as well as knowing the boundaries on freedom and is something more than mere imitation or recitation of rules. (6) There should be alternate forms of a good test to minimize the security problems of a testing program and to suit the specific needs of different organizations that might wish to use the test.

Implicit in these assumptions are some criticisms of the traditional methods of testing language proficiency. The better traditional tests (including TOEFL, the most recent product of the Modern Language Association) are, for the most part, composed of completion or multiple choice items which are scored right or wrong. Although a carefully constructed test of this type is generally more reliable and presumably more valid than an essay evaluation, the right-wrong scoring procedure inevitably equates the grammatically determined decision with the case of a slight stylistic preference. Imposing an ad hoc, subjective, weighting system to compensate for the initial error does not seem to be a satisfactory solution for the problem. Two sample items from the TOEFL Handbook for Candidates (ETS, 1963) will illustrate the point:

1. "This ballpoint pen won't write."
   "What's the matter ______ it?"
   (a) for (b) with (c) of (d) by   Answer (b)

2. "Because he had little education, his knowledge of the subject was ________."
(a) limited (b) small in quantity (c) minor
(d) not large at all Answer (a)

Although both of these items may index some important aspect of language, the "wrong" answers to item #2 are obviously not wrong in the same way, nor in the same degree, as those in item #1.

Secondly, the goal of a nationally standardized exam would seem to be in conflict with what is known about regional and specialist group differences in language usage (Malstom, 1959). In order to make national standardization somewhat meaningful, testers are pushed in the direction of formal written English (even in oral tests) even though they may be well aware that much of the S's communication will employ informal oral English or a dialect.

Thirdly, the difficulty and cost of constructing comparable multiple forms of the traditional type of test is prohibitive for all but large organizations that specialize in testing. Consequently, a small local organization which has need of a proficiency test must subscribe to the services of a testing organization or make up their own test, for which they have neither the time nor the necessary skilled personnel. Further, the time problems created by the security problems associated with the standardized exam prohibit the use of the best instruments. That is, a test which must be sent away for scoring, involving days or weeks of delay between testing and the reporting of scores, is of little value to a person who must make placement decisions in a matter of hours.

There can be little doubt about the need for tests of language proficiency. The increasing problems of college admission and placement, the increasing demand for training in foreign languages of all kinds and, most pressing, the increasing numbers of foreign students coming to our colleges and universities are all problems which demand reliable, valid, and practical tests of language proficiency as part of their solution. According to an Extract from Testing the English Proficiency of Foreign Students (distributed by Educational Testing Service, 1961), a conference on this subject was held in Washington, D.C. in 1961. This conference was jointly sponsored by the Modern Language Association of America, the Institute of International Education, and the National Association of Foreign Student Advisors. The following quotation is taken from the proceedings of that conference:

> The conference goes on record as recognizing the desirability of, and urgent need for a comprehensive program using carefully constructed tests of the English proficiency of foreign students, suitable and acceptable to all educational institutions in the United States and to various other organizations, chiefly governmental.

The conference recommended the development of a proficiency battery to measure (1) control of English structure, (2) auditory comprehension, (3) vocabulary, (4) reading comprehension, and (5) writing ability.

Such a battery was developed, the TOEFL referred to earlier. In addition, the conference recommended further research in language aptitude testing and measures of oral production skills.

The Test of English as a Foreign Language (TOEFL) has not been universally accepted. According to a service memorandum (ETS, 1965), 2,979 foreign students were tested between February of 1964 and May of 1965. A more recent communication from ETS indicates that 34,774 foreign applicants to U.S. colleges were tested between February of 1964 and April of 1967. Certainly a number of U.S. colleges that have a significant foreign student enrollment have not adopted TOEFL as an admission requirement. Perhaps this is due, at least in part, to some of the criticism outlined above.

Nevertheless, TOEFL is undoubtedly the most authoritative and the most complete test of English proficiency available today. For that reason, it was adopted as the primary criterion measure for assessing the validity of the CLOZENTROPY test developed in this study. The TOEFL and CLOZENTROPY tests were administered to a group of foreign students at the University of Colorado. Comparable reliability coefficients and a matrix of intercorrelations among the components of the two test batteries are given in the results section of this report.

A Mathematical Definition of Abnormality-Compatibility

A key element in the development of a new kind of language proficiency test was the development of procedures which permit an alternative to the right-wrong scoring system and to the subjective evaluation of essays. These procedures were derived from information theory.

Shannon and Weaver, in *The Mathematical Theory of Communication* (1949), provide us with a numerical expression for the *average absolute entropy* (H) of a source or code. With n independent symbols, each with a probability of occurrence p, the average amount of information (entropy) that can be transmitted with that symbol set is

$$-(p_1 \log_2 p_1 + p_2 \log_2 p_2 + p_3 \log_2 p_3 + \ldots + p_n \log_2 p_n) \text{ or } H =$$

$$-\Sigma\, p_i \log_2 p_i .$$

With two equally probable choices, H has a value of 1. With four equally probable choices, this expression has a value of 2. With four choices and the probabilities .37, .25, .21, and .17, the approximate value of H is 1.9381. The point, as far as this paper is concerned, is that one can describe a set of discrete, independent elements with a figure that expresses both the number of elements in the set and the probabilities of all the elements. The value (H), for any set of equally probable elements, increases as the number of elements increases and has a maximum value for any finite number of elements when they are equally probable.

A second component can be derived from the preceding discussion, but it is highlighted by Gleason (1961, p. 377). "The amount of information in any signal is the logarithm to the base two of the reciprocal of the probability of that signal. That is: $I = \text{Log}_2\, 1/p$." That value which was called the average absolute entropy, H, can now be seen to be the weighted average of the I values. $H - I$, then, is a deviation score (D) which expresses the extent to which a given symbol transmitted by a source carries more or less information (is more or less surprising) than the average of that source's transmissions.

If one thinks of a number of colored balls in an urn being drawn one by one (with replacement), I may be said to describe the "surprise value" of drawing any given color of ball from the urn. Its value will depend on the proportion of all the balls in the urn that are that color. H describes the average surprise value of drawings from the urn and will reflect the number of different colors represented in the set and the proportion of the total set which is of each color. D, then, represents the difference between the surprise on a given drawing and the average surprise of drawing. D will have a negative value when the drawing is more surprising than usual and a value of zero when the drawing is no more or less surprising than expected.

D can be taken as a measure of the "abnormality" of a given outcome of the system.

If, instead of an urn, we think of a context (such as a word association test) in which a unitary linguistic response is called for; and instead of colored balls, we think of the discrete linguistic responses of n independent subjects; then, by substituting the relative frequency of the *different* responses for p in the formulas above, we obtain a measure of the abnormality of an individual's response.

D scores from a number of items can be meaningfully added, and the composite score obtained in this fashion for an individual automatically takes into account the relative difficulty of the items as it is reflected in the amount of agreement among the members of the criterion group. In short, the D score would seem to have many of the desirable properties of the normal z, although it is derived from nominal rather than interval data.

D or *Sum D* can be computed for an individual with reference to a group of which he is a member, and it can also be computed with reference to an external criterion group (of which he is not a member). In this report, when D is computed for an "insider" it will be referred to as an "abnormality score." When D is computed for an "outsider" (with reference to an external criterion group), it will be called a "compatibility score."

Given an appropriate method of eliciting unitary linguistic responses from a group of people, D should be an appropriate measure of an individual's language proficiency with regard to that group. If the *sum* of D for an individual across a number of items is approximately zero, it would be interpreted that the individual's linguistic patterns are approximately normal. (It could also mean that there are no identifiable patterns or norms of language usage.) If the individual's score is positive and different from zero, it would presumably mean that he "conforms" to majority opinion in matters of linguistic choice. If an individual's *sum D* score, over a large number of items, is extreme and negative, it would indicate that there are norms of usage in the group to which he does not conform. A large negative score would, then, indicate an individual who would probably have difficulty understanding other members of the group and being understood by them.

The procedure which has just been described requires only the classification of responses into categories. It does not require making a priori judgments of the rightness or wrongness of a particular response. It takes into account the freedom of choice, or lack of it, exhibited by a group in a particular context.

Cloze Procedure

One way of eliciting unitary linguistic responses from groups of people in a well-controlled and efficient manner is cloze procedure. Cloze procedure was originated by W.L. Taylor in 1953, although it has antecedents in Gestalt psychology and in the common sentence completion technique. It was developed as an index of readability and was defined by Taylor (1954, p. 3) as "a psychological tool for gauging the degree of total correspondence between (1) the encoding habits of transmitters and (2) the decoding habits of receivers."

The procedure itself is quite simple. One selects a written passage and deletes every n^{th} word, replacing the deleted words with blanks of a uniform size. Subjects are then asked to replace the missing words to complete the passage. In Taylor's application to readability measurement, every fifth word was deleted because he found it made optimum use of the sampled material and allowed all sorts of words to be represented according to the proportion of their occurrence (1956, p. 48). Exact replacement of a missing word was counted as correct, and S's scores over fifty blanks seemed to give a sufficiently reliable ranking of different manuscripts. Cloze tests correlate highly with findings of the Dale-Chall and Flesch formulas on standard materials, but when these methods are applied to writing like that of Gertrude Stein, cloze scores seem to be a better measure of real difficulty (Taylor, 1953).

Taylor reports (1956) a study in which cloze scores were obtained along with comprehension scores from independently validated multiple choice tests and AFQT intelligence scores. Before and after cloze tests correlated .88, an indication of the reliability of the procedure. Cloze correlated with the comprehension test .80 and with AFQT .74. From Taylor's work, it seems to follow that cloze procedure can be used to measure learning, comprehension, intelligence, message difficulty, or any language-related variable depending on the application or way in which it is administered.

In his dissertation, Taylor (1954) also utilized the responses to cloze blanks to obtain an estimate of the number of different responses that subjects would supply to a given blank and the probability of each different response. From these, he calculated the entropy of a given blank (the H value referred to earlier). Assuming the deletion system to provide a representative sample of the message, he obtained estimates of the average entropy of the message or source. These figures correlated negatively with cloze scores (the proportion of exact replacements) across messages, but little else was done with this particular measure.

Darnell (1960) computed the entropy measure from cloze data and found significant differences among blanks and message treatments, where the treatment variable was the order of a set of fifteen sentences. He also suggested some minor changes in the computation procedure so that the entropy values obtained would not directly depend on the number of Ss employed. It was evident from this study that one can obtain reasonable estimates of the possible different responses to a given blank and, with a fairly large sample of persons, reasonable estimates of the probabilities of the different alternatives.

At least three groups of researchers have looked at cloze procedure as a possible measure of language proficiency. Carroll, Carton, and Wilds (1959) explored the possibility of using cloze procedure for this purpose and used three kinds of scoring systems. They tried giving credit for exact replacement only, for any grammatically correct response, and for the most frequent response. They rejected the immediate use of cloze procedure, because the measures which they obtained were relatively unreliable and too heavily influenced by such things as reasoning ability and ideational fluency. It must be pointed out that, in any case, they were using a right-wrong criterion that has little theoretical advantage over other types of proficiency tests. The use of the "most frequent response" system implicitly recognizes the problem but does not solve it.

Weaver and Kingston (1963) did a factor analysis of cloze scores and several other measures of language proficiency. Specifically, they used the Davis Reading Survey, five subtests of MLAT, the STEP listening test, three subtests of the Ohio State Psychological exam, and eight cloze tests. The eight cloze tests were combinations of two manuscripts (an essay and a speech), structural vs. lexical deletion systems, and silent reading by Ss vs. oral presentation by the experimenter. Three orthogonal factors were obtained and labeled "verbal comprehension," "redundancy utilization," and "rote memory--flexible retrieval." All but one of the cloze tests had their highest loadings on the redundancy utilization factor and none of the other tests load highest on this factor. The lexical deletion of the speech manuscript read aloud was the "odd" cloze test and had its highest loading on the verbal comprehension factor. The intercorrelations among cloze tests were about the same as those among other types of tests and generally higher than correlations between cloze and other tests. Weaver and Kingston suggest that cloze procedure is measuring something different than the other tests and that this is of interest. They also note that MLAT had its own unique factor in their analysis.

Weaver and Kingston (1963) also used the exact replacement method of scoring cloze procedure, and they note (p. 258) that this method does not take into account the fact that blanks differ in difficulty. The distinction which they made between structural (every n^{th}) deletion and lexical (only nouns and verbs) deletion systems was an attempt to deal with this difference. It is also interesting to note that they named the cloze factor "redundancy utilization" indicating cognizance of the relevance of information theory.

Holtzman and Hopf (1965) have also studied cloze procedure as a possible measure of language proficiency. Two manuscripts and exact replacement scoring were used. Differences were noted in the correlations between the two cloze messages and elements of the Holtzman-Spencer test battery given at the same time. Examination of the two manuscripts showed that every 10th word deletion system had eliminated a significantly larger proportion of function words and pronouns from one of the manuscripts. This finding would support the earlier suggestion that some weighting factor should be introduced to take account of the differences in difficulty of blanks.

Holtzman and Hopf conclude that cloze procedure does measure a significant element of language proficiency and that further research is justified. Their stated intent was, however, to turn from validation of cloze procedure as a measure of proficiency to a search for particular cloze texts which might be more successful than others.

Taylor (1953, p. 417) suggested, in contrasting his measure of readability with others, that "one may think of cloze procedure as throwing all potential readability influences in a pot, letting them interact, then sampling the result." Similarly, one can argue that as a measure of language proficiency, cloze procedure is a method of sampling the interactions of all the available influences. Each of the tests with which cloze scores have shown moderate correlations may be said to measure some rather specific knowledge or skill related to the use of language. The use of compound batteries of these tests argues that no one is an adequate predictor of success in the use of language. There is, of course, no evidence that cloze scores are more adequate than other kinds of tests as predictors of linguistic success. However, one can reason that an individual's available vocabulary, his knowledge of syntax, his awareness of cultural values associated with specific words or patterns, his sensitivity to numerous stylistic factors, and his integration of various contextual constraints will influence his choices of response to a mutilated cloze passage.

In natural language usage, the individual does make choices. Some of his choices are more influenced by some of the factors mentioned than others. It would seem more realistic, more appropriate, to examine the composite influence of all these language factors on an individual's linguistic behavior than to measure awareness of them in isolation and attempt to put them back together by some type of multiple regression equation. Acceptance of this proposition seems to be implied by the fact that many testing programs include a "speech" or "theme" in spite of the inherent subjectivity and low reliability of such measuring procedures. In the summary of conference decisions from the TOEFL conference mentioned earlier (ETS, 1961), the following statement appears:

> Writing ability is to be tested by objective techniques, not by the scoring of writing samples. However, an unscored composition will be furnished to test users for whatever use they may wish to make of it.

This statement seems to express a certain dissatisfaction with both objective testing and the available alternative.

CLOZENTROPY procedure is another alternative. The combination of cloze procedure and the entropy analysis described above would seem to have a number of advantages, both theoretical and practical, over other methods of measuring language proficiency. CLOZENTROPY procedure could be employed as follows.

First, a criterion group, consisting of 100 or more native speakers of a given language, would be selected to provide a reasonable standard of language usage. They should be representative of the group with whom the individual to be tested wants or needs to communicate. Then, a message in the target language with every n^{th} word deleted would be presented to this criterion group for completion. The number and frequency of the different responses to each blank would be tabulated and the average entropy (H) of each blank computed.

The same message test form would then be administered under similar conditions to a subject or group of subjects whose language proficiency is of interest. The S's response to each blank would be compared to the responses given by the criterion group to that same blank. If S's response occurred in the criterion group, it would be assigned a probability value equal to the proportion of the criterion group giving the same response. If S's response were a new response--one that had not occurred in the criterion group--it would be assigned a probability of $1/n$ (where n is the number of people in the criterion group). (Rather than assume that a response which did not occur in the criterion group is impossible--probability zero--the assumption is that S's response might have occurred if there had been one different person in the criterion group. This provides, at least, a usable estimate of the probability of the response and is giving S the benefit of the doubt.)

Given these assumptions, an I value can be computed for each response of each S. Given H and I for each blank, D can be computed for each subject and each blank, and sum D for a given subject is his test score. We have called this score a compatibility score, because it reflects the extent to which S's responses "stand out" in the array of responses from the criterion group.

The same kind of scores (sum D) can be computed for each member of the criterion group (in this case called an abnormality score), and a percentile rank in the criterion group provides a basis for interpreting S's proficiency score.

Since any other type of test would, ideally, require a standardizing group for interpretation of the scores, the procedure outlined above is not necessarily more difficult or expensive than other available procedures. It does utilize more information from the standardizing (criterion group) sample than normal standardizing procedures do. A second percentile ranking could be obtained from Ss tested over a period of time. (The percentile scores provided for interpretation of TOEFL scores are based on the foreign students tested to date.)

The extract from conference decisions from the MLA conference (ETS, 1961) recommends "Some native speakers of English should also be tested for initial information on difficulty, and as a check against faulty items, misleading directions, and the like." This aspect of the recommended validation procedures is built in the CLOZENTROPY procedure.

Pilot Studies

In the fall of 1965, at the University of Colorado, a series of pilot studies was begun to investigate the feasibility of CLOZENTROPY procedure. A passage from *Understanding Other Cultures* (Brown, 1963), selected for minimal cultural bias, was chosen as a test passage. The message, which dealt with fundamental likenesses in all cultures, was prepared for cloze procedure by deleting every fifth word. The instructions told *Ss* to write one word and only one word in each blank so that the completed passage would "sound like English."

This test form was then administered to 100 native speakers of English, predominantly freshman liberal arts majors, at the University of Colorado. Only the first fifty blanks were completed by all of this group in a fifty minute period, so only those fifty blanks were scored. The entropy analysis was performed on these data and abnormality scores obtained for each *S*.

For 72 of these *Ss*, scores on the Language Aptitude Test (LAT) were available from the University Placement Office. A product moment correlation was computed between the LAT scores and the *sum D* (abnormality) scores. The resulting coefficient was .70, providing limited affirmation of the validity of CLOZENTROPY procedure.

The same test form was also given to 22 foreign students enrolled in a remedial English course and *sum D* (compatibility scores) obtained for each foreign *S* in terms of the native speakers' responses. A ranking of these *Ss* on "general English proficiency" was obtained from the instructor of the course based on diagnostic examinations, course projects, and personal contact over a period of several weeks. A Kendall rank correlation was computed between the teacher's ranks and compatibility ranking. The resulting *tau* was .64 ($p < .001$). All of these *Ss* had been previously judged deficient in English, and it seems reasonable to infer that the correlation would have been somewhat higher had the sample included some *Ss* with adequate English proficiency.

In the fall of 1966, the same test form was administered as part of the Colorado placement battery to forty foreign students. Evaluation was based on the same criterion group data as the earlier study. Again, the sampling was somewhat selective, in that certain of the foreign students were judged "obviously proficient" and excused without taking the test battery. The test battery contained (1) the Gates Reading Survey, (2) the Lado Oral Comprehension Test, (3) a dictation test, (4) an oral interview, and (5) the CLOZENTROPY test. Tests 3 and 4 are of local origin and are subjectively evaluated on a 100 point scale taking into account grammar, punctuation, spelling, and "appropriateness of response to questions." A correlation analysis ($n = 40$) of the test battery produced the following matrix of correlations.

	1	2	3	4
1 (Gates)				
2 (Lado)	.54**			
3 (dictation)	.78**	.59**		
4 (interview)	.38**	.59**	.48**	
5 (CLOZENTROPY)	.61**	.34*	.63**	.23

(Note: * is significant at .05, ** is significant at .01)

This matrix shows that the CLOZENTROPY test correlated significantly with all but the oral interview, and that it correlated somewhat better with the reading and dictation tests than with the tests of oral ability. Given that two of these tests were scored subjectively, and the other two are scored on a right-wrong basis, these results were viewed as encouraging further research with the CLOZENTROPY procedure.

Another short pilot project, part of the regular evaluation procedures in one of the writer's courses, has relevance here. A passage from the assigned text was dittoed with every fifth word deleted and administered as a cloze test to 22 students. The responses were scored in two ways--by counting the number of exact replacements of deleted words and by the entropy procedure (scoring each individual against the total group). The cloze scores and the CLOZENTROPY scores (based on exactly the same data) correlated only .62 ($p < .01$). This finding suggested the possibility that research with the CLOZENTROPY procedure might produce a more favorable result than earlier research with cloze procedure as a measure of proficiency would indicate.

The study which is reported here is not, then, simply a rerun of past studies of cloze procedure. The scoring system which is used, as an addition to the established procedure, changes the whole complexion of the investigation. This study is not, simply, a validation study of another language proficiency test. Language proficiency is dealt with a kind of flexible conformity to the norms of a language community rather than the degree of rigid adherence to a particular model of language usage. The modified procedure (CLOZENTROPY) permits the investigation of the language community itself (as represented by the criterion group) as a determinant of the individual's ability to communicate with language. It permits an investigation of an individual's proficiency in a given content area relative to others who "know the language" but are equally uninformed in specific content and relative to others who are experts in both language and content. It also would permit (though it is beyond the scope of this study) the examination of the proficiency of college professors in using the English of the foreign or native student community.

Since the CLOZENTROPY procedure evaluates an individual against the background of a group's responses to a specific test passage, it is expected that obtained proficiency scores will be less influenced by the idiosyncracies of the author of the test passage, the specific content of the test passage, absolute difficulty of the test passage, chance deletion of certain types of words, and conditions of test administration (so long as they are constant for criterion group and test subjects) than would exact replacement scores.

Chapter II

Sampling and Testing

This study is an attempt to determine the reliability, validity, and practicality of a CLOZENTROPY test of the English language proficiency of foreign students. A CLOZENTROPY test was compared to the Test of English as a Foreign Language (TOEFL) currently being administered by Educational Testing Service. In addition, tests were performed to determine whether there are differences among categories of native speakers of English, among categories of foreign students, and among kinds of prose material used in the CLOZENTROPY test. This chapter will discuss the methods employed in sampling test material, native speakers of English, and foreign students. It will also describe the administration and scoring of the two tests.

The Test Passages

The samples of prose material used as test passages in this study were selected on the following criteria: (1) Passages should be representative of the kinds (subject matter) of material with which the Ss tested do, normally, have to contend. For example, students majoring in engineering should be tested on samples taken from lecture or text materials in current use in engineering courses. Students majoring in English literature should be tested on samples of literature used in courses they will have to take to reach their objectives. (2) Samples should represent the range of difficulty in materials the student is expected to use. Although the scoring procedure, theoretically, takes account of differences in difficulty among different form class deletions, it is entirely possible that a person with minimal proficiency will be increasingly disadvantaged as the difficulty of the material increases. That is, a more difficult passage might be expected to be more discriminating among levels of proficiency. Therefore, at least two samples from any given content area should be used--one rather easy and one rather difficult as determined by the Flesch readability index or other comparable measure. (3) The total sample should be long enough to get a reliable measure of an S's performance but not long enough to introduce a significant fatigue factor. (4) The sample should be relatively culture free. That is, samples from American history probably should not be used because of the presumed advantage that this would give the native American criterion Ss over foreign Ss. (5) The test material should not have been read, by either native or foreign Ss recently enough that rote memory could be expected to influence the individual's word choices.

Approximately fifty textbooks were examined with these criteria in mind. Students and faculty were consulted as to the representativeness of content and difficulty level of a variety of textbooks. These consultations revealed the fact that the topic of thermodynamics was common to all the divisions of engineering, and language or literature were common to most liberal arts specialties. Faculty members suggested books in these areas that had been used at the University of Colorado or were being considered for adoption.

Four textbooks were ultimately chosen which seemed to satisfy the major criteria. Within these four books, passages of continuous prose 500 words in length, of an appropriate difficulty level, were selected. The graduate engineering sample (GE) was taken from Heat and Thermodynamics by Mark W. Zemansky. The graduate liberal arts passage (GO) was selected from Philosophy of Language by William P. Alston. The undergraduate engineering sample (UE) was taken from Physics: For Students of Science and Engineering by David Halliday and Robert Resnick. The undergraduate liberal arts passage (UO) was taken from College English the First Year, edited by J. Hooper Wise et al. The four passages were found to have Flesch reading ease scores of 23, 24.9, 39.5, and 44.5 respectively. (Symbols in parentheses will identify the test passages as they appear in Appendix A.)

The Criterion Groups

Sincy the primary focus of this study was the testing of proficiency in English for foreign born college students, it was decided that the criterion groups should represent the competition that foreign students have to face. The criterion groups should, therefore, be made up of active students at an American university who are native users of English.

Since there are two major divisions of students with regard to level (graduates and undergraduates), it was decided that the criterion groups should represent these two divisions. Since the University of Colorado recognizes two distinct varieties of English (as do other universities) by maintaining two English language programs--one in the College of Engineering and one in the College of Arts and Sciences--it was decided to represent this distinction in the selection of the criterion groups. Therefore, it was decided to employ four criterion groups, graduates and undergraduates in engineering and graduates and undergraduates in non-engineering subjects. (The latter category is so named because there was no ready justification for excluding majors in the Schools of Business, Education, Journalism, etc. who utilize the A & S English language program.)

A list of foreign students was obtained from the foreign student office, a list of graduate students was obtained from the graduate school, and a student directory (listing all students) was readily available. (It was necessary to use all of these, because the student directory does not distinguish among majors at the graduate level, and the graduate list does not indicate home addresses.) Students on either list that were classified as having undetermined majors or as commuters were eliminated from the population. Undetermined majors were eliminated because we could not classify them, and commuters because we had no hope of gaining their cooperation in evening and weekend test sessions.

The graduate school list was then divided into two parts, engineering and non-engineering, and all names were eliminated which also appeared on the foreign student list. Each of the lists were then numbered systematically and a sample of fifty drawn at random from each list. Due to difficulties encountered later, some modifications in this procedure were necessary, but an attempt was made to obtain a random sample of native graduate students in each category.

To obtain a representative sample of undergraduates, the student directory (an alphabetical listing of all students registered at the University) was marked off in one inch segments, the segments numbered, and a sample of segments drawn at random. The first name in a randomly chosen segment that suited the specified criteria was assigned to one of the two undergraduate groups until fifty were obtained in each category. This procedure also had to be modified, but an attempt was made to obtain a representative sample of the undergraduate native population.

Letters were sent to these 200 native users of English asking them to participate in a research project and promising five dollars each for taking a test with a two hour time limit. A self-addressed postcard was enclosed asking them to select one of the available test dates. After one week, 73 affirmative replies and 24 rejections had been received. Of the 200 native students, better than fifty per cent didn't respond at all to the first request. Ultimately, with additional letters and personal phone calls, 110 of the original native sample were persuaded to participate. (Samples of the letters sent appear in Appendix B.)

The commitment of five dollars per person made over-sampling an unfeasible procedure. Instead, as rejections came in, the original sample was supplemented with additional "random" selections. Open testing hours from 8 a.m. to 8 p.m. were established for a period of one week. Finally, after a month, it became apparent that continuation of the described procedure would take an indefinitely long time to reach the goal of 200 native students. Realizing, too, that (after the first refusal) the "random" sample had changed to a sample of volunteers, we resorted to personal contacts and accepted unsolicited assistance from anyone who fitted one of the four categories to reach the objective of fifty in each category.

Examination of the available data gives no reason to suspect that the ninety "replacements" differed in any significant respect from the 110 from the original sample, but the fact remains that those who refused to participate were not available for comparison. It was extremely frustrating to watch our research ideals crumble on contact with reality, but in working with live human subjects, one is forced to take what he can get or abandon the research. We chose the former.

Foreign Student Ss

The population of foreign students on the Colorado University campus was identified by a list obtained from the foreign student office. This list indicated the student's major and level of study. The names were divided into four sets on the basis of this information, each set was numbered, and random samples of twenty were drawn from each of the four categories--graduate engineers, graduate non-engineers, undergraduate engineers, and undergraduate non-engineers.

A letter (with an enclosed reply card) was sent to each foreign student in the sample (See Appendix C). The letter asked each student to take the TOEFL and an experimental language test (a commitment of about six hours time) and promised to pay five dollars on completion of both tests. Fifty-one foreign students responded--32 agreed to participate, seventeen refused, and two made ambiguous responses.

Replacements for the seventeen refusals were selected by matching on level of study, academic major, native country, and sex. New letters were prepared, with what we thought were stronger motive appeals, asking them to consider the five dollars as a gift rather than payment for services rendered. These letters were sent to the matched replacements and a supplemental sample chosen at random. A continuous effort was made to contact those who had not responded. Finally, as the date of the special administration of TOEFL approached, we attempted to call all those who had not responded as well as the rest of the population in the undergraduate engineering category.

On the day of the TOEFL administration, we had positive commitments from 69 foreign Ss with approximately equal division in the four categories. Forty-eight persons appeared to take the TOEFL--ten undergraduate engineers, eleven undergraduate non-engineers, thirteen graduate non-engineers, and fourteen graduate engineers. Two of the foreign students were native speakers of English but gave their place of birth and residence as outside the United States. One person who took the test was not technically a foreign student, but a recent emigrant whose native language was Japanese.

Since one of the major objectives of this study was to compare TOEFL with the CLOZENTROPY test, no further attempt was made to obtain more foreign Ss. Instead, we devoted our energy to getting these 48 Ss to take the second test. Although the sample was not as large as intended and not random as intended, it was the best we could do without scheduling a second administration of TOEFL, which would have meant a significant delay in the testing program. Examination of the Ss obtained gave no reason to suspect that they were in any way atypical of the foreign student population, but, of course, those who refused were not available for comparison.

Test Administration

The TOEFL was given to 48 Ss on Saturday, April 13, 1968. It was administered in a large, well-lighted, lecture hall with ample room to permit alternate seating. This special administration was carried out with complete conformity to ETS's restrictions. It was supervised by a staff member from the University of Colorado testing center with three assistant proctors. The total time required for the administration of TOEFL was approximately 4-1/2 hours.

Two testing sessions were scheduled for the CLOZENTROPY test--the Wednesday evening and Saturday morning following the TOEFL administration. These test sessions were held in the same room as the TOEFL administration and under very similar conditions. Foreign Ss and native Ss were mixed together in these administrations. A total of 95 native Ss and 44 foreign Ss took the test in these two sessions. All Ss were able to complete the test in the two hour time limit.

The CLOZENTROPY test booklets were stenciled with one page of instructions, including examples. The order of the four test passages was varied so that all possible orders occurred with equal frequency. The booklets were distributed to Ss at random to control for any possible order effects in taking the test. (A sample booklet appears in Appendix A.)

Since 105 native students and four foreign Ss had not been tested at the conclusion of these scheduled sessions, open testing hours and individual appointments were arranged in another classroom to complete the testing program. Since the written instructions appeared to be quite adequate and the two hour time limit ample, the changes in administrative procedure did not seem to cause any serious complications.

Fourteen days elapsed between the administration of the TOEFL and the last foreign student's taking the CLOZENTROPY test. There were 28 days between the first and last administration of the CLOZENTROPY test to native Ss.

The TOEFL was, of course, scored by ETS and the scores for each S reported to the project director. The CLOZENTROPY test was scored locally. This was accomplished by punching the responses, in alphabetical form, on data punch cards, six responses per card. Then, through the use of the CDC 6400 computer and a program written by Dr. David R. Saunders of the CU Psychology Department, the responses to each item were tabulated and values for H (for each item), I and D (for each response), and sum D (for each S) computed. Each native S was scored against the total group of native Ss to obtain an "abnormality score." Foreign Ss were scored against each of the four criterion groups separately and against the total criterion group to obtain "compatibility scores." The computer program provided an alphabetical listing of the different responses that were given to each item, the frequency of each response, and the D value for each response. It also calculated the sum D for each S and transformed these scores to make the distribution of criterion group scores approximately normal with a mean of 1.00 and a standard deviation of approximately .30. (Samples of the computer output for the scoring program appear in Appendix D.)

Two kinds of analyses were planned for these data. First, we wanted to do a correlation comparison of the TOEFL and CLOZENTROPY tests. Secondly, we planned to do analyses of variance of the CLOZENTROPY data for both native and foreign Ss to determine the effects of each of the variables involved in the CLOZENTROPY test. Detailed descriptions of the ANOVA designs appear in the next chapter along with the results of the total analysis.

Chapter III

Results

The major purposes of this study were to determine the reliability, validity, and practicality of a CLOZENTROPY test for measuring English language proficiency. To those ends, the battery of four CLOZENTROPY test passages was administered to 48 foreign students along with the Test of English as a Foreign Language (TOEFL) which is currently administered by the Educational Testing Service.

Reliability

To assess the reliability of the CLOZENTROPY battery, a Hoyt reliability coefficient was computed. Following Kerlinger (1965, pp. 432-440), the method involves a two-way analysis of variance, items by subjects. For this analysis, the four subtests were treated as items. The reliability coefficient equals one minus the ratio of error variance to individual variance

$$(r_{tt} = 1 - {}^{V}e/V_i).$$

The obtained coefficient for the total CLOZENTROPY test was .859. For comparison purposes, the same kind of analysis was performed on the TOEFL scores, treating the five subtests as items. The total TOEFL reliability coefficient, obtained from the same 48 Ss and by the same method, was .864. Since both coefficients would round to .86, as far as can be determined from these data, the reliabilities are the same, and they are satisfactorily high. These analyses utilized the BMD program 02V. Reliability coefficients were not computed for individual subtests in either battery.

Validity

One basis for determining the validity of a measuring instrument is to correlate the results of that instrument with another accepted measure of the variable under study. Scores from the CLOZENTROPY and TOEFL test batteries, along with grade point averages for the 48 foreign Ss, were submitted to a regression-correlation analysis (BMD 02R). Due to the fact that graduates and undergraduates are graded on different scales, three analyses were actually performed, one for the total group and separate analyses for graduate and undergraduates.

As it turned out, little could be learned from the step-wise regression analysis except that no combination of elements from the two test batteries would account for more than one third of the variance in GPA for the total group (53 per cent for graduate students). For that reason, the regression analysis is omitted from this report. The correlation matrices that were obtained are displayed in Table 1.

Table 1

Correlations between CLOZENTROPY, TOEFL, and GPA

(1) Difficult Engineering
(2) Diff. Non-Engineering
(3) Easy Engineering
(4) Easy Non-Engineering
(5) Total CLOZENTROPY
(6) Grade Point Ave.
(7) Listening Comp.
(8) English Structure
(9) Vocabulary
(10) Reading Comp.
(11) Writing Ability
(12) Total TOEFL

Correlation Matrix for Total Group N = 48

	2	3	4	5	6	7	8	9	10	11	12
1	.621	.654	.592	.846	.177	.648	.470	.478	.508	.520	.625
2		.533	.634	.840	-.086	.643	.668	.654	.494	.700	.771
3			.625	.839	-.148	.573	.508	.654	.529	.534	.686
4				.834	-.214	.625	.577	.658	.522	.605	.730
5					-.077	.736	.665	.733	.614	.700	.838
6						-.075	-.253	-.161	-.056	-.069	-.158
7							.590	.512	.639	.514	.753
8								.694	.404	.782	.856
9									.514	.789	.899
10										.410	.682
11											.877

Correlation Matrix for Graduates N = 27

	2	3	4	5	6	7	8	9	10	11	12
1	.637	.769	.478	.851	.448	.612	.580	.549	.444	.618	.686
2		.586	.609	.876	.161	.736	.698	.675	.543	.675	.815
3			.515	.849	.349	.525	.422	.697	.420	.607	.679
4				.756	-.122	.622	.577	.725	.415	.603	.743
5					.251	.750	.687	.794	.558	.741	.877
6						.150	.301	.159	-.026	.205	.205
7							.589	.523	.425	.595	.741
8								.628	.377	.811	.848
9									.480	.741	.883
10										.437	.629
11											.903

Correlation Matrix for Undergraduates N = 21

	2	3	4	5	6	7	8	9	10	11	12
1	.615	.568	.685	.853	.248	.676	.420	.399	.564	.409	.572
2		.447	.651	.794	-.002	.523	.606	.585	.468	.713	.690
3			.674	.819	-.129	.585	.530	.583	.642	.437	.659
4				.891	-.094	.606	.559	.564	.628	.587	.695
5					.005	.709	.633	.638	.688	.635	.780
6						-.016	-.073	-.109	-.086	-.033	-.080
7							.625	.469	.836	.405	.759
8								.771	.571	.809	.892
9									.584	.838	.906
10										.396	.785
11											.843

There are a number of things to be learned from observation of these matrices. For instance, the correlations between the total CLOZENTROPY scores and the total TOEFL scores are .838 (for the total group), .877 (for the graduate students), and .780 (for the undergraduates). These coefficients are highly significant (beyond .001) and indicate that approximately seventy per cent of the variance in the test scores is common to the two tests. Considering the structural and theoretical differences in the two tests, any larger correlation would be difficult to believe.

A second point of interest is the correlation between the listening comprehension subtest (variable #7) with the CLOZENTROPY test (#5) and the TOEFL score (#12) of which it is a part. For the total group, "listening comprehension" correlates .736 with CLOZENTROPY and .753 with TOEFL. For graduate Ss, the respective correlations are .750 and .741, and for undergraduates, .709 and .759. These correlations indicate that the CLOZENTROPY does as good a job measuring listening comprehension as does the TOEFL which specifically includes this subtest for that purpose. Correlations between CLOZENTROPY and other subtests of TOEFL are all highly significant, but listening comprehension is the only instance in which the CLOZENTROPY correlation actually exceeds the correlation with TOEFL. This is doubly surprising since the CLOZENTROPY test might, on a priori grounds, be thought to have more in common with the English Structure or the Reading Comprehension subtests. The pilot study given in Chapter I of this report would, also, have suggested a different result.

Another interesting finding is the lack of correlation between all the language proficiency tests and grade point average. The only correlation that is significantly different from zero is the Difficult Engineering test in the CLOZENTROPY battery. All other correlations with this measure of academic success are close to zero and tend to be negative, especially for the undergraduate Ss.

The correlations of each of the CLOZENTROPY subtests with the total TOEFL (.625, .771, .686, and .730) are very similar to the .70 found in pilot research between another fifty item CLOZENTROPY test and LAT scores.

The intercorrelations among CLOZENTROPY passages (based on the total group) range from .533 for the two "difficult" passages to .654 for the two engineering messages. For graduate students, the correlations range from .515 for the two easy messages to .769 for the two engineering passages. For undergraduates, the low correlation is .447 ($p < .05$) between the easy engineering and the easy "other" passage and the high is .685 between the difficult engineering and the easy "other" passage. These intercorrelations are additional indications of the reliability of the CLOZENTROPY procedure.

Effects of Six Independent Variables on CLOZENTROPY Scores

To determine the effects of the several variables involved in this kind of test, two analyses of variance were performed. A five-way analysis was performed on the data provided by native users of English who served as criterion groups in this study and a seven-way analysis of the foreign student data. Because it is somewhat easier to explain, the design and result of the native student analysis will be considered first. Both analyses used the BMD 08V computer program which handles nested designs.

ANOVA of Native Data. The 200 native Ss were scored against themselves so that each S's scores (four per S) reflect his standing in the total group. This factorial design for repeated measures can be described as a 2 x 2 x 50 x 2 x 2--two levels, graduate (L1) and undergraduate (L2), by two majors, engineering (M1) and other (M2), by subjects (nested within levels and majors), by two kinds of content, engineering (C1) and other (C2), by two levels of difficulty, easy (D1) and difficult (D2).

Table 2

Analysis of Variance of Native Student Data

Source of Variation	df	Mean Square	Error Term	F	p
Level (L)	1	4.2009	S(LM)	14.59	<.01
Major (M)	1	1.1276	S(LM)	3.91	<.05
Content (C)	1	0.0819	SC(LM)	1.63	n.s.
Difficulty (D)	1	0.0883	SD(LM)	1.93	n.s.
L x C	1	0.9005	SC(LM)	17.93	<.01
M x C	1	1.5727	SC(LM)	31.32	<.01
L x M x C	1	0.2163	SC(LM)	4.30	<.05
L x C x D	1	0.2980	SCD(LM)	7.06	<.01
Subjects [S(LM)]	196	0.2879			

Note: Interactions not significant at .05 and error terms with no intrinsic interest are omitted.

It is apparent from Table 2 that there is a significant difference between graduates and undergraduates in performance on this test. The level (L) effects produce an F of 14.59 with a chance probability of less than .01. Observation of the cell means indicates that this difference is in favor of the graduate students. (Significantly different cell means from this analysis are shown in Appendix E.)

There is a significant difference between engineers and non-engineers (F = 3.91, $p < .05$), and the engineers, as a group, score slightly higher.

There are no main effects of the content and difficulty variables, but this is an artifact of the scoring procedure, since the means of the four subtests are necessarily the same except for computational and rounding error. There are, however, interactions which involve these variables indicating that they are relevant to the overall analysis.

There is a significant interaction between level and content (see Figure 1). This interaction indicates that difference between graduates (L1) and undergraduates (L2) is greater on engineering content (C1).

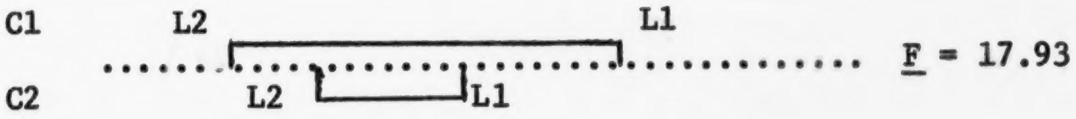

Figure 1. Level by content interaction

There is a significant interaction between major and content (see Figure 2). This interaction indicates that engineers are far superior on engineering material, while no real difference exists between engineers and non-engineers on non-engineering material.

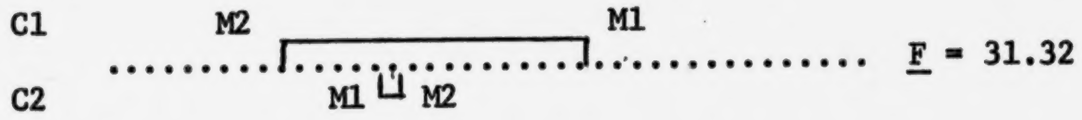

Figure 2. Major by content interaction

A significant three-way interaction among level, major and content (see Figure 3) is, perhaps, best described by noting the relatively high score made by graduate engineers on engineering material and the relatively low score made by undergraduate engineers on non-engineering material. One can also note that undergraduate engineers scored higher than non-engineers on both kinds of content. The simple fact of this three-way interaction underscores the relevance of these three variables to a study of language proficiency.

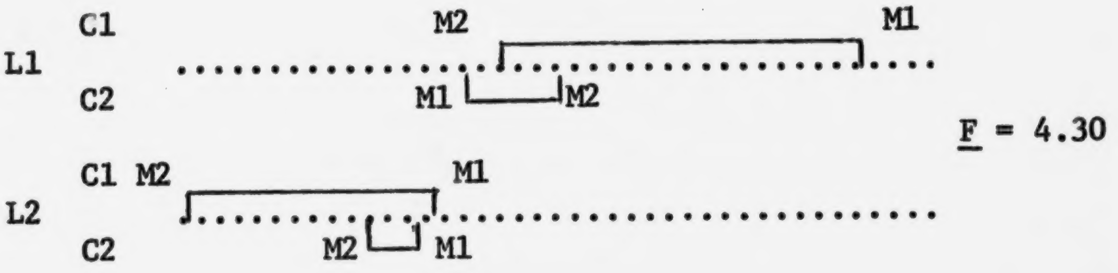

Figure 3. Major by level by content interaction

The difficulty variable, which did not have a significant main effect, does interact with level and content (see Figure 4). For both levels, the effects of content were greater in the more difficult material, but the graduates did better on engineering content and the undergraduates did better on non-engineering material.

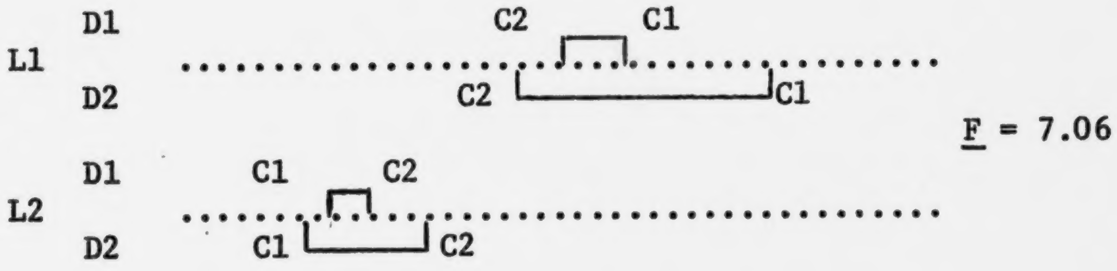

Figure 4. Content by difficulty by level interaction

ANOVA of Foreign Data. The analysis of variance of the foreign student data was very similar to the one just described. The number of subjects was smaller and there were two additional variables in this analysis. Although the test was administered to 48 Ss, the proposed analysis required equal numbers in each cell, so eight Ss were randomly deleted from the larger sets leaving ten Ss in each of the four categories. The additional variables were generated by scoring the foreign Ss against each of the four categories of native students separately. The two new variables were labeled "R" for rank of the criterion group (graduate and undergraduate) and "G" for major of the criterion group (engineering and other). Other variables and labels are the same as in the preceding analysis, except that we are now dealing with foreign students.

The seven-way analysis of variance (repeated measures, nested design) can be described as a 2 x 2 x 10 x 2 x 2 x 2 x 2--levels (graduate and undergraduate) by major (engineering and non-engineering) by subjects (nested within the two preceding variables) by content (engineering and other) by difficulty (easy and difficult) by rank of the criterion group (graduate and undergraduate) by major of the criterion group (engineering and non-engineering). Results of this analysis are summarized in Table 3.

Table 3

Analysis of Variance of Foreign Student Data

Source of Variation	df	Mean Square	Error Term	F	p
Level (L)	1	2.6047	S(LM)	3.61	n.s.
Major (M)	1	0.0580	S(LM)	.08	n.s.
Content (C)	1	0.3267	SC(LM)	6.06	<.05
Difficulty (D)	1	0.2973	SD(LM)	3.14	n.s.
Rank (R)	1	0.4965	SR(LM)	215.87	<.01
Group (G)	1	1.7211	SG(LM)	452.92	<.01
M x C	1	1.5683	SC(LM)	29.01	<.01
C x D	1	0.6248	SCD(LM)	9.16	<.01
M x R	1	0.0224	SR(LM)	9.73	<.01
C x R	1	0.4216	SCR(LM)	248.00	<.01
C x G	1	0.1991	SCG(LM)	104.79	<.01
D x G	1	0.0444	SDG(LM)	44.40	<.01
M x C x R	1	0.0127	SCR(LM)	7.47	<.05
C x D x R	1	0.0095	SCDR(LM)	7.31	<.05
C x D x G	1	0.0727	SCDG(LM)	72.70	<.01
C x R x G	1	0.0619	SCRG(LM)	77.37	<.01
D x R x G	1	0.0348	SDRG(LM)	69.60	<.01
M x C x D x R	1	0.0072	SCDR(LM)	5.54	<.05
C x D x R x G	1	0.0239	SCDRG(LM)	29.87	<.01
Subjects [S(LM)]	36	0.7204			

Note: Interactions not significant at .05 and error terms with no intrinsic interest are omitted.

In contrast with the native student analysis, there is no significant difference indicated between graduates and undergraduates or between engineers and non-engineers among the foreign Ss. Although the variable, major, enters into several significant interactions, these differences are apparently not as great among foreign students as among native student Ss.

There is a significant difference (F = 6.06, p < .05) between engineering and non-engineering content with the mean score being higher on the engineering content. (Significantly different means from this analysis are shown in Appendix F.)

There is not a significant main effect on the difficulty variable, and such an effect is not precluded by the scoring system as it was in the native student analysis. Difficulty does, however, enter into significant interactions, indicating that it is a relevant variable.

It does make a difference which criterion group foreign *Ss* are scored against. Significant main effects are indicated for both rank of the criterion group ($F = 215.87$, $p < .01$) and major of the criterion group ($F = 452.92$, $p < .01$). Foreign *Ss* compare most favorably with (seem to be most similar to) undergraduates and non-engineers.

There is a significant interaction between content and major. As indicated in Figure 5, both majors do better with their own kind of content, but content makes more difference to engineers.

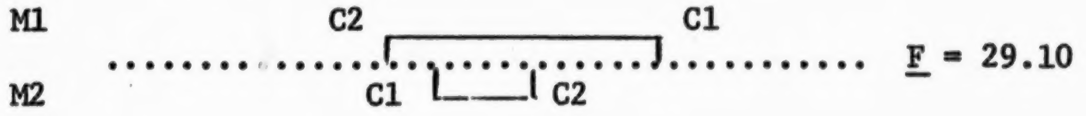

Figure 5. Content by major interaction

The significant interaction between content and difficulty displayed in Figure 6 indicates that content differences are greater in the easier material. The preference is for engineering material at the easy level, but there is no real preference at the difficult level.

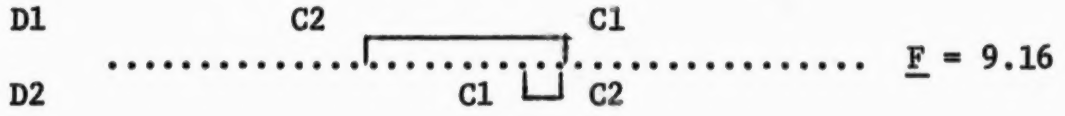

Figure 6. Content by difficulty interaction

The significant interaction between content and rank of the criterion groups displayed in Figure 7 indicates that content effects are almost entirely attributable to the undergraduates in the criterion groups, or, it might be better to say, that there are no content effects when foreign *Ss* are scored against graduate students.

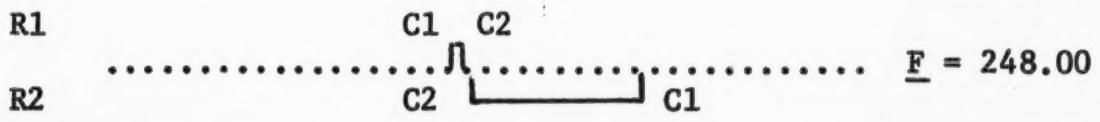

Figure 7. Content by rank of criterion group interaction

Figure 8 displays the interaction between major and rank of the criterion group. Although engineers tend to score higher than non-engineers, this difference is greater--though means are lower--when foreign *Ss* are scored against graduate students.

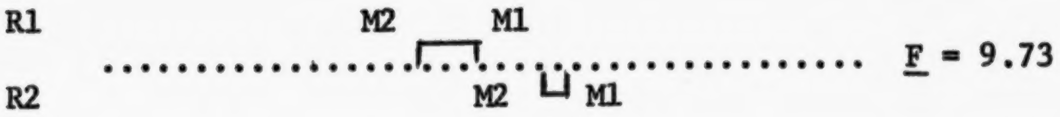

Figure 8. Major by rank of criterion group interaction

The major of the criterion group interacts with the content variable, as shown in Figure 9, and with the difficulty variable as in Figure 10. Both content and difficulty tend to have greater effects when foreign Ss are scored against non-engineering native English users. So amplified, the means show a relatively higher performance for engineering material and the more difficult material. It is important to consider here that there is a kind of interaction between "nationality" and other variables built into the scoring system, a kind of mirror image effect. The foreign Ss tend to do better, comparatively speaking, where the native Ss do poorly.

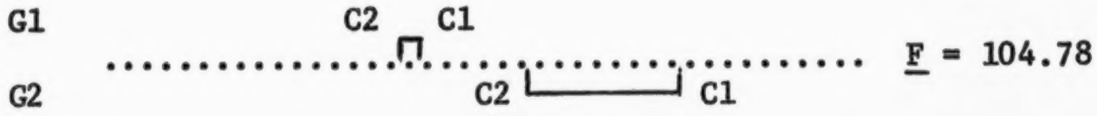

Figure 9. Content by major of criterion group interaction

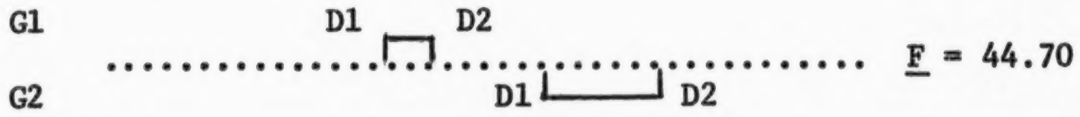

Figure 10. Difficulty by major of criterion group interaction

It is probably already clear that "language proficiency" is a complex phenomenon, but there is further evidence of the interrelatedness of five of the six variables dealt with in this study.

A significant three-way interaction among major, rank of the criterion group, and message content is illustrated in Figure 11. The association between major and content is apparent, but it diminishes to near zero when non-engineering Ss are scored against undergraduates.

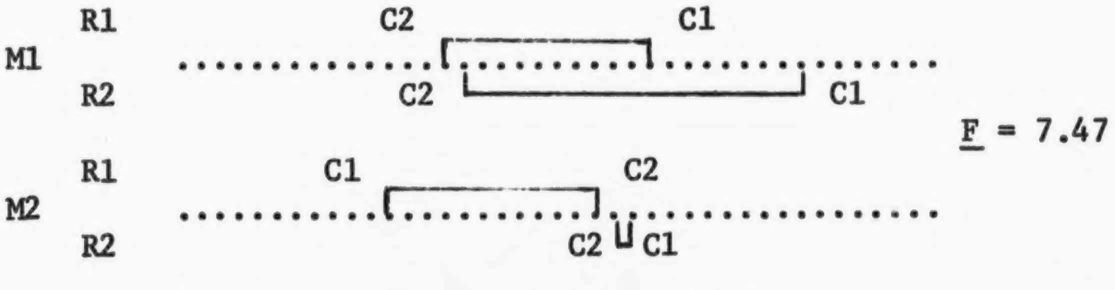

Figure 11. Content by major by rank of criterion group interaction

A three-way interaction among content, difficulty, and rank of the criterion group is shown in Figure 12. The effects of the difficulty variable are larger in non-engineering content, and for both kinds of content, tend to be larger when foreign Ss are compared to graduate students.

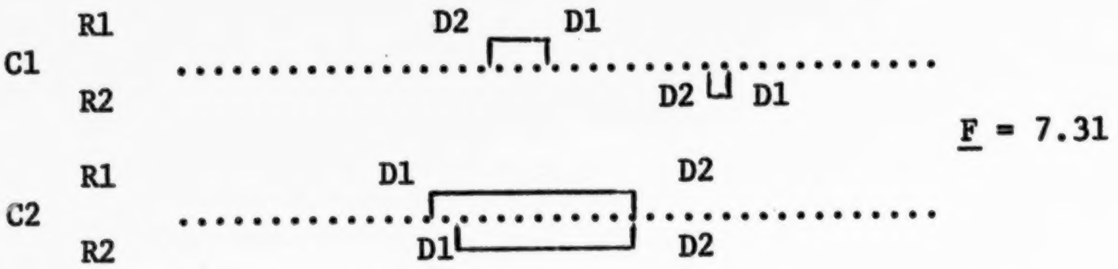

Figure 12. Difficulty by content by rank of criterion group interaction

There is a significant three-way interaction among content, difficulty, and major of the criterion group shown in Figure 13. In engineering content, the effects of difficulty are rather small and in the opposite directions for the two criterion groups. In non-engineering content, the differences in difficulty level are larger, and the difficult messages get higher scores regardless of the major of the criterion groups.

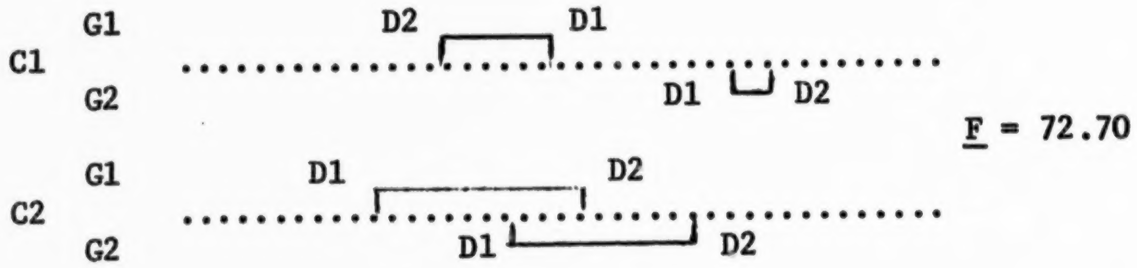

Figure 13. Difficulty by content by major of criterion group interaction

The three-way interaction among content, rank of criterion group, and major of the criterion group is shown in Figure 14. It indicates that, in engineering content, scored against engineers, Ss do better when compared to undergraduates. On non-engineering content scored against engineers, Ss do slightly better compared to graduates. Scored against non-engineers, Ss do better on both kinds of content when compared to undergraduates.

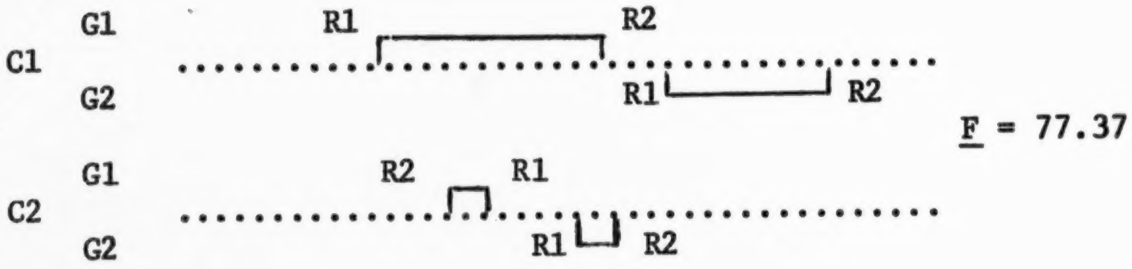

Figure 14. Content by rank of criterion group by major of criterion group interaction

The last significant three-way interaction is displayed in Figure 15. This interaction is among rank of the criterion group, major of the criterion group, and difficulty of the test passage. Rank of the criterion group makes more difference on easy material when the criterion group is made up of engineers, but on difficult material, rank makes more difference when the criterion group is also restricted to non-engineers.

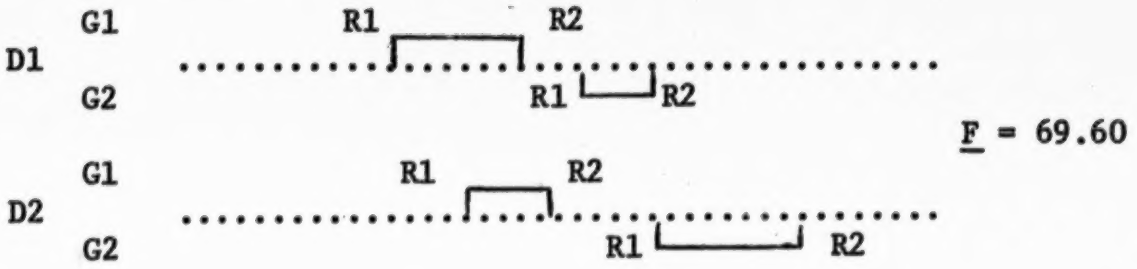

Figure 15. Difficulty by rank of criterion group by major of criterion group interaction

There are two significant four-way interactions displayed in Figures 16 and 17. The first is among major of the Ss, content of the test material, rank of the criterion group, and difficulty of the test material.

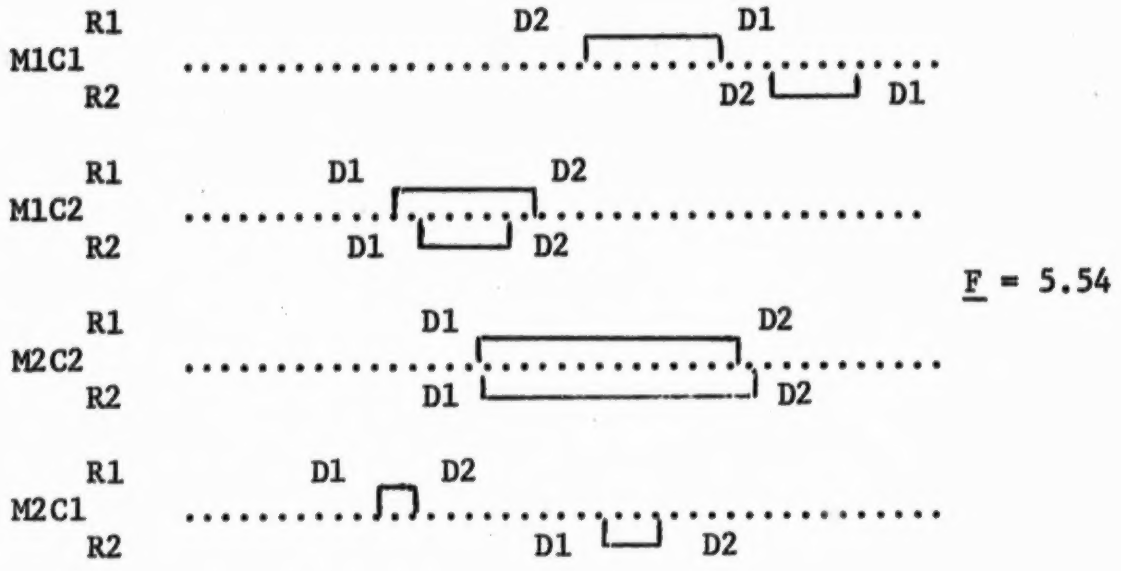

Figure 16. Major by content by difficulty by rank of criterion group interaction

The second four-way interaction is among content of the test material, difficulty of the test material, rank of the criterion group, and major of the criterion group.

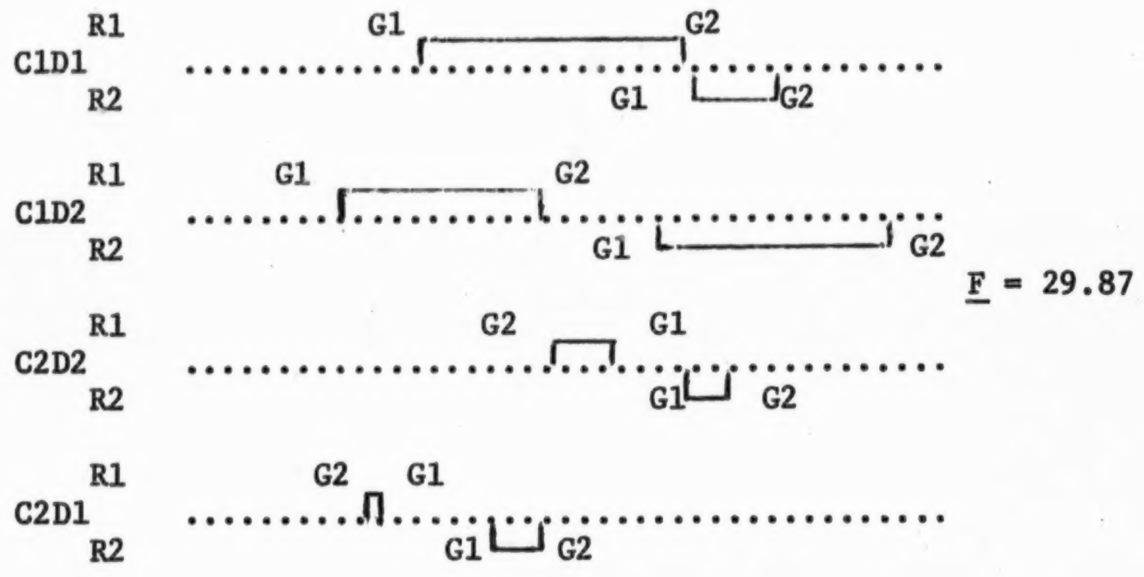

Figure 17. Content by difficulty by rank of criterion group by major of criterion group interaction

Such complex interactions defy verbal description. Hopefully, their visual presentation will contribute to an impression of the complexity of the language proficiency concept. The only variable which did not contribute to any of the significant effects is rank or level of the foreign student Ss. It is also apparent from viewing Figures 16 and 17 that had this study been restricted to easy, non-engineering content scored against graduate students and the major of the foreign Ss disregarded (as we at one time considered doing), the result would have been much less informative.

Although no test of significance was performed to determine whether the foreign Ss differ from the native Ss, the difference is quite apparent. The scoring procedure for the native students with a transformation in the direction of normality makes the native mean on each subtest approximately 1.00 (the largest deviation from this was 1.09 on the difficult engineering message) with a standard deviation of approximately .30. By comparison, the foreign student mean on the difficult engineering message was .625 with a standard deviation of .296; on the difficult non-engineering message the mean was .642 and the standard deviation was .321; on the easy engineering message the mean was .644 and the s.d. .328; on the easy other message the mean was .537 and the s.d. .278. The most plausible explanation for these differences is that foreign Ss do tend to make responses that are "unusual" compared to the array of responses provided by native users of English.

Summary

In the ANOVA treatment of foreign student CLOZENTROPY data, content of test material, difficulty of test material, rank of the criterion group, major of the criterion group, and major of the Ss were shown to have significant main or interaction effects. No significant differences are attributable to the graduate-undergraduate distinction among foreign students.

In the ANOVA treatment of native student CLOZENTROPY data, level, major, content of test material, and difficulty of test material were all shown to have significant main or interaction effects. Since the native Ss were scored against the total group, the criterion group variables did not enter into that analysis.

In the correlation analysis of the foreign student data, the correlation between the total CLOZENTROPY test and the total TOEFL was .838 (n = 48). This correlation, considering only the 27 graduate students, was .877 and for the 21 undergraduates .780. The correlations with grade point average, for both CLOZENTROPY and TOEFL, were, for all practical purposes, zero with a consistent tendency toward negative relationships. A particularly interesting finding was the correlation between the CLOZENTROPY test and the Oral Comprehension subtest of the TOEFL battery. This correlation was .736, which compares very favorably with the correlation between the total TOEFL and the subtest, .753.

Hoyt reliability coefficients were computed for both of the test batteries treating the subtests as items. The CLOZENTROPY coefficient was .859. The TOEFL coefficient, based on the same 48 Ss and the same computational method, was .864. As far as can be determined from the responses of these Ss, there is no difference in the reliability of the two test batteries, and it is quite high in both cases.

Chapter IV

Summary and Conclusions

Two tests were administered in this study and a comparative analysis made of the results. One of the tests was the Test of English as a Foreign Language (TOEFL), a battery made up of five subtests of the multiple choice type. The five subtests are said to measure (1) oral comprehension, (2) English structure, (3) vocabulary, (4) reading comprehension, and (5) writing ability. A "total score" is also provided which is twice the actual sum of the subtest scores. This test is regularly administered by Educational Testing Service of Princeton, New Jersey at testing centers around the world. In this case, ETS granted permission for a special administration for research purposes at the University of Colorado. The administration of the test was supervised by a representative of the University's testing service according to guidelines provided by ETS. The answer sheets were returned to ETS for scoring. Forty-eight foreign Ss took this test at a cost to the project of ten dollars per subject.

The second test was an experimental battery (CLOZENTROPY). Evaluation of the CLOZENTROPY battery, in terms of reliability, validity, and practicality, was the central focus of this study. CLOZENTROPY procedure, as the name is intended to suggest, is a combination of cloze procedure (the data collection instrument) and a method of analysis derived from information theory (an entropy analysis). All tests can be described as a combination of a data collection instrument and a method of scoring. Most tests place the emphasis on the data collection instrument and utilize a rather simple "right-wrong" scoring system. The value of such tests depends almost entirely on the skillful construction of items and the acceptability of the criteria for making the right-wrong judgment on each item. CLOZENTROPY procedure, on the contrary, utilizes an extremely simple method of constructing test items and a rather complex, mathematically precise, scoring system which avoids entirely the right-wrong judgment on an item by item basis.

To create a data collection instrument of the cloze type, one selects a sample of prose material and replaces every $\underline{n}^{th}$ word with a blank. Ss are then instructed to replace the missing words with single words that fit the context. In this study, four samples of textbook prose were used representing different kinds of content and different levels of difficulty. Each sample was approximately 500 words in length. Every 10^{th} word was deleted and replaced by a ten-space blank. Two of the samples were, according to the Flesch formula, rather easy and two rather difficult. Two represented engineering content and two liberal arts content. These four test forms were the subtests of the CLOZENTROPY battery.

The entropy analysis of the data obtained from cloze procedure is as follows: Considering the array of responses from some specific group of Ss to a particular cloze item, determine the number of different responses and the relative frequency of each. Assuming that the relative frequency is a good estimate of the probability of each different response, calculate the "average surprise value" of responses to that item ($\underline{H} = \Sigma\ p_i \log_2 p_i$).

(This value may be called the *entropy* of the blank, and it is a measure of the freedom-of-choice available to respondents.) Calculate the "surprise value" or "information value" of each of the different responses ($I = \log_2 1/p$). Obtain the difference (*D*) between the *I* value for each response and the *H* value for each blank ($H - I = D$). Repeat this procedure for each item and sum the *D* scores for each *S* across all items in the test. The *sum D* score is an indication of the extent to which the individual tends to give responses that are more or less unusual in the context of the group's responses, so it has been called an "abnormality score" when an *S* is scored against a group of which he is a member.

One may also score "outsiders" against a particular criterion group (as was done in the case of the foreign students in this study) by assigning the outsider's response the relative frequency of the same response among members of the criterion (insiders) group. (In this case, "compatibility score" seems a better name for *sum D*.) In the event that an outsider emits a unique response (that did not occur in the criterion group) or any *S* omits an item, one may assign a minimum probability estimate of $1/n$ (where *n* is the number of respondents in the criterion group).

This procedure has the effect of a built-in item analysis. For example, if every *S* emits the same response, the *D* is zero. If every *S* emits a different response, all *D*s are zero. If a variety of responses occur, and they are not equally frequent, *D* reflects the relative popularity of each response, the more popular response having a higher value. If an *S* emits a response that differs from a response on which all other *S*s agree, his *D* score is maximum for that item and negative.

In this study, 200 native users of English served as criterion groups for the evaluation of 48 foreign students. Abnormality scores were computed for each native *S* and compatibility scores for each foreign *S* as measures of *S*'s proficiency in the use of English.

The regression-correlation comparison of the two tests and the ANOVA treatments of native and foreign CLOZENTROPY data vindicated every hope of the designers of the experimental test.

The finding of almost identical reliability coefficients for the TOEFL and CLOZENTROPY batteries (both round to .86) seems to be strong support for the reliability of CLOZENTROPY. Although the coefficients are not quite as high as those reported for TOEFL based on repeated measurements and much larger samples, they were calculated by identical procedures and are based on the same 48 *S*s. Consideration should also be given to the fact that the CLOZENTROPY battery, although it contained 200 items, required less than half as much time as the TOEFL to administer. One could, presumably, increase the reliability of the CLOZENTROPY battery by increasing its length. The potential increase in power would not, however, seem to justify the added burden on test subjects.

The correlation obtained between CLOZENTROPY and TOEFL of .780 (for undergraduate *S*s), .877 (for graduates), and .833 (for the total group) seems rather strong support or the validity of the CLOZENTROPY battery. The two tests would appear to be measuring, for all practical purposes, the same thing. According to the reliability coefficients, approximately fourteen per cent of the variance in each set of test scores is error variance. Approximately seventy per cent of the total variance is common to the two tests, so the communality accounts for almost all the reliable variance in either test. To the extent, then, that TOEFL is an acceptable measure of English proficiency, the CLOZENTROPY battery must also be acceptable.

From another point of view, neither test seems to be a valid predictor of academic achievement. The correlations with grade point average (for foreign students) were pratically zero with tendencies in the negative direction. Since it is difficult to imagine how cognitive development could be independent of proficiency in the code which prevails in the learning environment, it seems almost more plausible to assume that cognitive development and academic achievement are unrelated. It could be that teachers tend to compensate for the student's weakness in the language. It could also be that a majority of the students in this sample are performing at a level above that critical level of proficiency necessary for normal academic achievement.

Given the extreme differences in conceptualization and form of the two tests studied here, the evidence for comparability can be taken in two ways. If one accepts the assumptions about language stated in the first chapter of this report, one must admire the skill of the creaters of TOEFL who managed to reflect the norms of the college student community while operating in a highly prescriptive framework. This, of course, may mean that the college student population has so thoroughly internalized the model on which TOEFL is based that the suspected discrepancy between the "ideal" and the "actual" language does not exist for this population. One would certainly expect larger differences in the results of the two tests if the criterion groups had not been so thoroughly trained in formal English usage. If, on the other hand, one accepts the concept of "standard English" and compliance with the rules of standard English as a better criterion for assessing proficiency, one must marvel at the extent to which unskilled test makers, employing the CLOZENTROPY procedure and substituting the behavior of a selected criterion group for judgmental skill, have apparently succeeded in reaching the same objective.

In analysis of native CLOZENTROPY data, it was found that every variable involved in the study (level of student, major of student, content of test material, and difficulty of test message) had some kind of main or interaction effect on the resulting test scores. These results suggest that there are several different kinds of language proficiency and that the general proficiency test may not be optimally sensitive to the specialized needs of a particular student.

In the analysis of foreign student CLOZENTROPY data, the variable, level of the student, produced no discernible effect. Major of the student and difficulty of the message had only interaction effects. Content of the message and the two criterion group variables (rank and major) had both simple and interaction effects.

The apparent homogeneity of the foreign student sample is of particular interest. The fact, visible in the interaction patterns, that foreign students seem to use English most like that of the undergraduate non-engineer, may have some bearing. A language teaching program must start somewhere, and one could guess that it typically starts with conversational English and/or with the study of simple English literature. One could also guess that no distinction is typically made between graduates and undergraduates in English classes or between engineers and non-engineers, on the assumption that basic English is basic English. Only one set of norms is provided with the TOEFL scores, so, presumably, the same standard is commonly used for admission to graduate work as for admission to undergraduate work.

The differences found in the native groups would suggest that the optimal language program for the individual foreign student would take the level and major variables into account. In the present system, engineers and graduate students are probably operating at a disadvantage.

The development of specialized tests of the CLOZENTROPY type, utilizing appropriate content and criterion groups of native English peers, would probably influence programs in English for foreign students in a desirable direction. At the same time, such tests would provide the foreign student with a more relevant criterion for self evaluation. It is easy to understand why the foreign student who is eager to obtain a graduate degree in engineering is not overly enthusiastic about required courses in "appreciation of simplified literature." If he were convinced, however, that his classmates are getting more than he is out of material that he wants to understand, and if English courses were available which promised to do something about that, he might participate with higher motivation and with greater profit to all concerned. The logic of the CLOZENTROPY procedure should be of some value in obtaining such motivational conviction.

CLOZENTROPY has several practical advantages over the more common kind of test. One, which has already been mentioned, is that its administration requires less than half the time required by TOEFL with almost identical results. Secondly, the creation of alternate forms requires merely the selection of a new sample of material and administration to a criterion group to obtain a scoring key. The criteria used in this study for the selection of test material were so general one would expect other samples to produce extremely similar results. The intercorrelations among the four samples of material used in this test would tend to support the repeatability of this result.

The ease with which alternate forms of the CLOZENTROPY test can be created leads to an additional advantage. That is, the security problems of the standardized test are practically eliminated. It is conceivable that new forms could be developed for every major administration. Certainly, such a large number could be developed in such a short time that effective "cheating" would be a very time consuming task. Since there would be no need to use old items on new test forms, and answers are automatically validated by the scoring procedure and criterion group, the loss of a test form would only be an inconvenience. The individual institution could also develop its own test forms and provide its own criterion group data, thereby eliminating the possibility of outside contamination.

A fourth advantage which CLOZENTROPY has over many conventional tests is ease of administration. The continuity of the task, the simplicity of instructions, and the minimal security problem make this a very easy test to administer. Only normal proctoring is necessary. It may be administered in a large group or to individuals in a clinical setting. Since answers are written in and order of the test forms may be varied, even neighborly copying is severely inhibited.

A fifth, and very important, advantage derives from the fact that the CLOZENTROPY procedure allows for a free kind of response rather than selection from a limited set of alternatives. It, therefore, measures encoding skill as well as decoding skill, at least in a limited way. Also, the free response form reduces the likelihood that an individual will score high by chance alone. An S may score below his potential on this kind of test if his attitude toward the test form or the content of a particular passage should interfere with his thought processes, but there is no logical way that he can profit from making purely random responses. This is particularly important if the test is used for screening purposes. An S who is required to take remedial work because he does not, on a particular occasion, give an adequate representation of his true ability, may be released from that requirement when new evidence points out the error, but an error of the other kind is not typically discovered until it's too late to do anything about it.

One major limitation of the CLOZENTROPY test is the scoring procedure. It would be practically impossible to tabulate the responses and compute the abnormality or compatibility scores for any reasonable number of subjects or items without computer assistance. Even with the scoring key, it is immensely difficult to look up each response on an alphabetical list, write down the corresponding D value, and accumulate totals over a number of items. Most organizations that might wish to use such a test have access to adequate computer hardware, but some adaptation of available software would be necessary before the system could be fully utilized. If further research bears out the results of this study, such adaptations will undoubtedly occur.

Although the results of this study can be applied strictly only to these subjects and messages and to those hypothetical populations of which they are representative, there is considerable justification for believing the results can be replicated in a wide range of circumstances. Neither subjects nor messages were selected on any criterion unique to this situation. At the time of this writing, no accidental biases have been discovered. The most likely contaminant is the self selection factor in the sample of subjects. Of potential Ss selected at random, that subset which refused to participate might reasonably contain a higher proportion of those who doubted their own linguistic ability than the set which did participate in the test program. Even this doesn't seem very likely, because repeated attempts were made to reassure them that tests, rather than subjects, were being evaluated, and the excuses which were offered seemed, for the most part, to be legitimate and unrelated to the variables being studied.

The CLOZENTROPY procedure is definitely not restricted to this application. It would seem to be applicable to the problem of testing the "linguistic mobility" of minority group members in relation to the larger society. By obtaining samples of the language actually used in some particular social group and samples of people who have been successful (by whatever criterion) in the use of that language, this procedure could be used to determine the compatibility of a non-member's language patterns with those of the group. In that kind of application, the usual standards of formal English seem much less appropriate than in the prediction of success in the academic community. One could even use minority group members who have "succeeded" as a realistic standard for aspiring members of that minority group. By simply changing the criterion group, the standard of evaluation is also changed.

Further research is planned which will apply CLOZENTROPY procedure to oral language. Although this written instrument was found to correlate rather highly with one measure of oral comprehension, there are reasons to pursue the issue. Given the requirement of almost instantaneous comprehension in oral discourse, and assuming that native speakers of English have greater freedom of choice in oral than in written English, this procedure may prove to be even more sensitive to deficiencies in oral English. Such investigation seems worthwhile if only to obtain comparative measures of oral and written language skills with comparable instruments.

One question that has been consistently raised by the few critics of CLOZENTROPY procedure is, "How do you tell the difference between the unusual response that is creative from the unusual response that is simply odd?" No satisfactory answer has yet been obtained, but, at least, it's a reasonable question in the context of this test procedure. Further investigations may very well bring to light a way of making this distinction. A measure of linguistic creativity would be very useful. At this point, however, the distinction does not seem to be a critical one, because excessive creativity would seem to interfere with the communication function of language as much as any other "abnormality."

Finally, it seems important to emphasize the fact that the mathematical procedure developed here is not tied to this particular data collection instrument. Cloze procedure came into being as a measure of readability. The entropy analysis, derived from information theory, may be employed with any data collection instrument in the analysis of any behavior, so long as behavioral norms are appropriate criteria for evaluation. It is almost coincidence that the two were combined in this way to produce what seems to be a reliable, valid, and extremely practical measure of language proficiency.

REFERENCES

Alston, William P., Philosophy of Language. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1964, pp. 66-67.

BMD Biomedical Computer Programs, Los Angeles: University of California, Health Sciences Computing Facility, Revised September, 1965.

Broadhurst, A.R. and D.K. Darnell, "Introduction to Cybernetics and Information Theory," Quarterly Journal of Speech, LI (December, 1965), 440-453.

Carroll, J.B., A.S. Carton, and Claudia P. Wilds, "An Investigation of 'Cloze' Items in the Measurement of Achievement in Foreign Languages," College Entrance Examination Board Research and Development Report, Laboratory for Research in Instruction, Harvard University, Cambridge, Massachusetts, April, 1959.

Darnell, D.K., "The Relation between Sentence Order and Comprehension of Written English," Unpublished M.A. thesis, Michigan State University, 1960.

_____________, "The Relation between Sentence Order and Comprehension," Speech Monographs, XXX (June, 1963), 97-100.

ETS (Educational Testing Service), Extract from Testing the English Proficiency of Foreign Students, Center for Applied Linguistics of the Modern Language Association of America, Washington, D.C., 1961.

___, Test of English as a Foreign Language, Handbook for Candidates, 1961.

___, "Memorandum Transmitting Scores on the Test of English as a Foreign Language," (mimeo) November, 1965.

Gleason, H.A., An Introduction to Descriptive Linguistics. New York: Holt, Rinehart, and Winston, Inc., 1965.

Halliday, David and Robert Resnick, Physics: For Students of Science and Engineering. New York: John Wiley & Sons, Inc., 1962, pp. 532-533.

Holtzman, P.D. and T.S. Hopf (Pennsylvania State University), "Cloze Procedure as a Test of English Language Proficiency," Paper presented to the Speech Association of America, December, 1965.

Kerlinger, F.N., Foundations of Behavioral Research. New York: Holt, Rinehart, and Winston, Inc., 1964.

Malstrom, Jean, "Linguistic Atlas Findings Versus Textbook Pronouncements on Current American Usage," English Journal, XLVII (April, 1959), 191-198.

Shannon, C.E. and Warren Weaver, The Mathematical Theory of Communication. Urbana: The University of Illinois Press, 1949.

Taylor, W.L., "A New Tool for Measuring Readability," Journalism Quarterly, XXX (Fall, 1953), 415-433.

___________, "Application of 'Cloze' and Entropy Measures to the Study of Contextual Constraints in Samples of Continuous Prose," Unpublished Ph.D. dissertation, The University of Illinois, 1954.

___________, "Recent Developments in the Use of Cloze Procedure," Journalism Quarterly, XXXIII (Winter, 1956), 42-48.

___________, "'Cloze' Readability Scores as Indices of Individual Differences in Comprehension and Aptitude," Journal of Applied Psychology, XLI (1957), 19-26.

Weaver, W.W. and A.J. Kingston, "A Factor Analysis of the Cloze Procedure and Other Measures of Reading and Language Ability," Journal of Communication, XII (December, 1963), 252-261.

Wise, J.H., J.E. Congleton, A.C. Morris, and J.C. Hodges, (eds.) College English the First Year. New York: Harcourt, Brace, & Company, 1956, pp. 235-236.

Zemansky, M.W., Heat and Thermodynamics. New York: McGraw-Hill Book Company, 1957, pp. 157-158.

Appendix A

A Sample Copy of the Test Instrument

Name ______________________________

Native Language ____________________

Number of years in the United States __________

Number of years experience with English ___________

A Test of English Language Proficiency

This booklet constitutes an experimental test of English language proficiency. It contains 4 passages from textbooks, each approximately 500 words in length. From each passage 50 words have been systematically deleted, so missing words may be of any syntactical type (nouns, verbs, articles, prepositions, etc.).

One word, and only one word, has been deleted wherever you find a 10 - space blank. (In one case a number has been deleted.)

You are not expected to fill in every blank with the exact word that has been left out. Your task is to fill in each blank with a word which "seems to fit" in the context of the passage.

You will be allowed a maximum of 2 hours to complete the entire test, so you should spend no more than 30 minutes on each passage.

We suggest that you scan an entire passage, go over it again filling in the "easy" blanks, then go back a third time and fill in the "difficult" ones. If nothing seems to fit in a particular blank, GUESS.

Do your best to fill in all the blanks.

SAMPLE: We _____1_____ that you scan an entire passage, go over it

_____2_____ filling in the "easy" blanks, then go back a

_____3_____ time and fill in the "difficult" ones. If

nothing _____4_____ to fit in a particular blank, GUESS.

Blank number 1 may be filled with "suggest" or "hope." Number 2 will permit "again" or once." Number 3 will allow "second" or "third," and the last one would allow "seems," "appears," "happens," or even "emerges."

ARE THERE ANY QUESTIONS?

(This passage is taken from a philosophy text)

Empiricist criteria of the sort we are considering are usually stated as genetic theories about the way people learn what words mean or the way words acquire meaning. This is, in part, a reflection of the fact that in British empiricism of the seventeenth and eighteenth centuries, epistomology and semantics were not really separated from psychology.

1 The separation is by no __________ complete today, but now we
2 are all well aware of __________ dangers of seeking answers to
3 questions of fact, including __________ fact, by the traditional
4 armchair methods of philosophy -- reflection __________ clarification.
5 If we really want to find out how __________ learn the meanings of
6 words and what mechanisms are __________ in such learning, there
7 is no substitute for careful __________ of the process itself; it
8 is ill-advised to __________ theories about this on *a priori*
9 considerations, such as __________ have in the preceding arguments.
10 Fortunately, it is not __________ to give these criteria a
11 genetic form. In general, __________ is possible to replace
12 any empiricist genetic account with __________ parallel statement
13 of what must be the case for __________ expression to have **a meaning**
14 Thus in place of __________ Lockean genetic account, we can
15 propose the following: in __________ for an expression to be
16 meaningful in my current __________ of it, it is necessary that
17 there be a __________ for the word to elicit in me
18 a certain __________ and vice versa. The formulation in
19 terms of ostensive __________ seems to be more wedded
20 to the genetic form, __________ it can be restated without
21 losing its empiricist force: __________ word can have a
22 meaning for someone only if __________ is able to pick out
23 its "referent" in his __________. This means that we have
24 shifted from the genetic __________ that a word has acquired
25 its meaning by way __________ an ostensive definition to the

26 requirement that it be __________ to give an ostensive definition.
27 Since genetic formulations are __________ easily convertible,
28 I shall continue to make use of __________ for the sake of easy
29 intelligibility. (The first argument __________ given for an empiricist
30 criterion, which as stated, supports __________ genetic criterion,
31 could also be reformulated along similar lines.) __________ begin to
32 emerge when we note that it cannot __________ the case that every mean-
33 ingful expression in the language __________ its meaning through direct
34 confrontation with an experienced referent. __________ account seems
35 plausible for common nouns denoting observable physical __________ --
"tree," "house," "cloud," adjectives connoting directly observable
36 properties -- "blue," "__________," "shiny;" and verbs that are concerned
37 with directly observable __________ -- "walk," "speak," "wave." However,
38 there are many other words __________ to these grammatical classes,
39 whose meaningfulness would not be __________ by any but the most hardy
40 empiricists and which __________ not possibly get hooked up with
41 their extralinguistic objects __________ this way, because the kind
42 of thing, property, or __________ involved is not directly observable. I
43 am thinking of __________ words as "society," "conscientious," "intelligent,"
44 "neurosis," "language," "education," "brilliant," "__________," "pray,"
45 "prosper." One cannot teach someone what the word "__________" means by
46 pointing to someone prospering while uttering the __________ in the way one
47 can teach someone what "run" __________ by (repeatedly) pointing to someone
48 running while uttering the __________. Of course, one can observe
49 instances in these cases. __________ can watch someone praying or
50 (engaged in) managing a __________, one can see a neurotic or an intelligent
or a conscientious person and can even observe him doing something that
displays his intelligence or conscientiousness or is a symptom of his neurosis.

U0-1

(This passage is taken from an English text)

Biography, which is simply the story of a man's life, has taken many forms. Today we think of a biography as being an entirely factual prose narrative. But the word has not always been so precisely used.

1 The __________ biographers were probably those minstrels
2 who praised the deeds __________ heroes in extemporaneous song.
3 As a hero's exploits increased--__________ fact or in imagination--
4 the songs grew in length __________ number, until at last the total
5 reached epic proportions. __________, for instance, was the genesis
6 of our most ancient __________, the Iliad and the Odyssey. Almost
7 as ancient, and __________ speaking more biographical, are the
8 accounts of the lives __________ the prophets in the Old Testament.
9 The earliest example __________ the carefully wrought, formal
10 type of biography was Plutarch's __________ Lives, written in the first
11 century A. D., which presented __________ a series of contrasting
12 pairs, the lives of a __________ of famous Greeks and Romans.
13 Plutarch's chief concern was __________ men as types of moral ex-
14 cellence or moral weakness, __________ than as individuals endowed
with complex personal characteristics.
15 During __________ centuries of development between Plutarch's
16 Lives and the biography __________ we know it today, biographical
17 writing took many forms, __________ miraculous lives of the saints and
18 a multitude of __________ accounts of kings and conquerors. The compelling
19 motives in __________ biographical endeavors were largely didactic or
20 commemorative. In the __________ and seventeenth centuries, however,
21 a greater degree of curiosity __________ how other people thought and
22 lived was reflected in __________ experiments in biographical writing
23 as prefatory biographical essays, character __________, printed funeral
24 sermons, letters and diaries. In short, emphasis __________ placed on
25 the historical rather than the ethical motive, __________ more attention

26 was given to the individual as a __________ being.

27 The next stage in the development of biography __________ around two
28 prominent eighteenth-century writers--Samuel Johnson and __________ Boswell.
29 Johnson insisted that the whole truth be given __________ a man:
30 "If a man is to write A __________, he may keep vices out of sight;
31 but if __________ professes to write A Life, he must write
32 it __________ it was." James Boswell followed this dictum
33 of his __________, and in consequence established for
34 biography a permanent place __________ a type of literature.
35 In The Life of Samuel __________, one of the most fascinating
36 books in all of __________ literature, Boswell captured Johnson's
37 individuality, wit, wisdom, and arrogance. __________ a work of art,
38 Boswell's masterpiece possesses charm, realism, __________ analysis,
39 and force far surpassing that of any of __________ predecessors.

40 The spirit of scientific accuracy and the respect __________ exhaustive
scholarly research that characterized the late nineteenth
41 century __________ in long, detailed biographies of the
42 "life and times" __________. These not only recounted the
43 events of a man's __________, but used that life as a
44 center around which __________ organize a history of the
45 times in which the __________ lived. As a consequence,
46 nineteenth-century biographies are of __________ scope. Lockhart's
47 life of his father-in-law, Sir __________ Scott, was first published
48 in ten volumes; Froude's Life __________ Carlyle, in four volumes;
49 Forster's Life of Dickens, in __________ volumes.

Among modern readers biographical writing competes
50 keenly in __________ with other types of literature. Recent biographers
have drawn heavily from modern psychology for method in character analysis and from drama and the short story for techniques of presenting their subjects.

(This passage is taken from an engineering text)

Carnot Cycle. During a part of the cycle performed by the working substance in an engine, some heat is absorbed from a hot reservoir; during another part of the cycle a smaller amount of heat is rejected to a cooler reservoir. The engine is therefore said to operate between

1 these two reservoirs. Since it is __________ fact of experience that

2 some heat is always rejected __________ the cooler reservoir, the

3 efficiency of an actual engine __________ never 100 percent. If we

4 assume that we __________ at our disposal two reservoirs at a

5 given temperature, it __________ important to answer the following

6 questions: (1) What is __________ maximum efficiency that can be

7 achieved by an engine __________ between these two reservoirs?

8 (2) What are the characteristics __________ such an engine?

9 (3) Of what effect is the __________ of the working substance?

10 The importance of these questions __________ recognized by Nicolas

11 Leonard Sadi Carnot, a brilliant young __________ engineer who, in the

12 year 1824 before the first __________ of thermodynamics was firmly es-

13 tablished, described in a paper __________ "Sur la puissance motrice

14 du feu" an ideal engine __________ in a particularly simple cycle

15 known today as the __________ Cycle.

16 In describing and explaining the behavior of this __________ engine,

17 Carnot made use of three terms: feu, chaleur, __________ calorique. By

18 feu he meant fire or flame, and __________ the word is so translated no

19 misconceptions arise. Carnot __________, however, no definitions for

20 chaleur and calorique, but in __________ footnote stated that they

21 had the same meaning. If __________ of these words are translated

22 as heat, then Carnot's __________ is contrary to the first law of

23 thermodynamics. There __________, however, some evidence that,

24 in spite of the unfortunate __________, Carnot did not mean the

25 same thing by chaleur __________ calorique. Carnot used chaleur

26 when referring to heat in __________, but when referring to the
27 motive power of heat __________ is brought about when heat
28 enters at high temperature __________ leaves at low temperature,
29 he used the expression chute __________ calorique, never chute
30 de chaleur. It is the opinion __________ a few scientists that
31 Carnot had in the back __________ his mind the concept of entropy
32 for which he __________ the term calorique. This seems incredible,
33 and yet it __________ a remarkable circumstance that, if the
34 expression chute de __________ is translated "fall of entropy,"
35 many of the objections __________ Carnot's work raised by Kelvin,
36 Clapeyron, Clausius, and others __________ no longer valid. In
37 spite of possible mistranslations, Kelvin __________ the importance
38 of Carnot's ideas and put them in __________ form in which they
appear today.

39 A Carnot cycle __________ a set of processes that can
40 be performed by __________ thermodynamic system whatever, whether
41 chemical, electrical, magnetic, in thermal __________ with a cold
42 reservoir at the temperature θ_2. Four __________ are then performed
43 in the following order: (1) A __________ adiabatic process is per-
44 formed in such a direction that __________ temperature rises to that
45 of the hotter reservoir, θ_1. __________ The working substance is
46 maintained in contact with the __________ at θ_1, and a reversible
47 isothermal process is performed __________ such a direction and to
48 such an extent that __________ Q_1 is absorbed from the reservoir.
49 (3) A reversible __________ process is performed in a direction
50 opposite to (1) __________ the temperature drops to that of the

cooler reservoir, θ_2. (4) The working substance is maintained in contact with the reservoir at θ_2, and a reversible isothermal process is performed in a direction opposite to (2) until the working substance is in its initial state. During this process, heat Q_2 is rejected to the cold reservoir.

(This passage is taken from a physics text)

Let us first consider a system in thermodynamic equilibrium. A system will be in thermodynamic equilibrium when it meets the following

1 requirements. (a) The system is __________ a state of mechanical
2 equilibrium -- there is no unbalanced __________ in the interior of the
3 system and no unbalanced __________ between the system and its
4 surroundings. (b) The system __________ in thermal equilibrium -- all
5 parts of the system are __________ the same temperature and this
6 temperature is the same __________ that of the environment. (c) The
7 system is in __________ equilibrium -- it does not tend to undergo
8 a spontaneous __________ of internal structure. A system in thermo-
9 dynamic equilibrium can __________ specified macroscopically by
10 giving the values of only a __________ quantities, such as pressure,
11 volume, temperature, and quantity of __________ particular substance.
12 Now suppose that we change the state __________ the system. A change
13 in state must involve some __________ from thermodynamic equilibrium.
14 For example, suppose that we change __________ system from one
15 state to another having just half __________ volume. Imagine that
16 we do this by quickly pushing __________ a piston. The system will
17 not be in thermodynamic __________: there will be relative motion
18 of its parts owing __________ unbalanced forces; temperature differ-
19 ences may set in because the __________ effects of the compression
20 may affect different portions of __________ system in different ways;
21 there may be chemical changes __________ changes in phase, such as
22 condensation. Of course, eventually, __________ left to itself, the
23 system may reach a new __________ of thermodynamic equilibrium. During
24 the process of change, however, __________ equilibrium does not exist.
25 Most processes of interest can __________ thought of as beginning in an

26 equilibrium state, passing __________ nonequilibrium states, and ending in
27 another equilibrium state. Thermodynamics __________ to understand such pro-
28 cesses. But rather than concerning itself __________ the details of the highly
29 complex processes whereby nonequilibrium __________ approach equilibrium,
30 thermodynamics seeks instead to obtain simple and __________ information about
31 such processes by comparing their behavior to __________ of an ideal process,
32 called a reversible process. In __________ reversible process we change
33 the state of a system __________ a continuous succession of equilibrium states.
34 For example, suppose __________ try to reduce the volume
35 of a system to __________ its original value by a succession
36 of small changes, __________ first increase the force on the
37 piston by a __________ small amount. This will reduce the
38 volume of the __________ a little; the system will depart
39 from equilibrium, but __________ slightly. In a short time
40 the system will reach __________ new equilibrium state.
41 Then we increase the force on __________ piston again by a very
42 small amount, reducing the __________ further. Again we wait for
43 a new equilibrium state __________ be established, and so forth.
44 Hence, by many repetitions __________ this procedure we finally
45 achieve the required change in __________. During this entire
46 process the system is never in __________ state differing much from
47 an equilibrium state. If we __________ carrying out this procedure with
48 still smaller successive increases __________ pressure, the intermediate
49 states will depart from equilibrium even __________. By indefinitely increas-
50 ing the number of changes and correspondingly __________ the size of each
change, we arrive at an ideal process in which the system passes through a
continuous succession of equilibrium states.

Appendix B

Sample Letters to Native Students

UNIVERSITY OF COLORADO
DEPARTMENT OF SPEECH AND DRAMA
BOULDER, COLORADO 80302

March , 1968

Dear

Would you be willing to help us make an important decision? Would you like to receive $5.00 during the week of April 15th? Are you interested enough in the English language or the problems of testing language proficiency to spend a couple of hours taking a test? If your answer to any of these questions is "yes", read on.

We are currently engaged in a research project, sponsored by the U. S. Office of Education, to validate a new kind of test for measuring the English proficiency of foreign students. Part of the problem is to discover the kind of English used by native American students, as a basis for evaluation of the foreign student's English.

This is where you come in. You have been chosen in a random sample of college students from C. U. to provide a standard of measurement. If English is your native language, it is very important that you participate, because every substitution we are forced to make will qualify the results of the total study.

The test we are asking you to take will be given Wednesday evening, April 17th and Saturday morning, April 20th. Please indicate on the enclosed card which of these times you would rather take the test. If you absolutely can't help us out, please return the card indicating that fact. If you agree to help, we'll send you a reminder card giving the exact time and place of the test administration.

On completion of the test, we'll hand over the $5.00.

This may be your only opportunity to take a test on which any answer you give is, by definition, a right answer.

Sincerely yours,

Dr. Donald Darnell, Project Director
Associate Professor
Department of Speech and Drama

DEPARTMENT OF SPEECH AND DRAMA
UNIVERSITY OF COLORADO
BOULDER, COLORADO

April 12, 1968

Dear

This letter disproves the old adage that opportunity knocks but once.

We do need your help in the validation study of a new test of English proficiency, and the test dates are rapidly approaching. So, we are reminding you of your opportunities to

a) make a contribution to the body of scientific knowledge about language,

b) help establish a reasonable standard of acceptable English usage, and

c) receive $5.00 in cash.

The test will take only two hours. It will be given in Guggenheim 201 on Wednesday, April 17th (7:00 to 9:00 p.m.) and on Saturday, April 20th (9:00 to 11:00 a.m.). You may choose either one of these times.

Please inform us of your interest.

Sincerely yours,

Dr. Donald K. Darnell, Project Director
Department of Speech and Drama
Extension 7488

p.s. It's a paper and pencil test--you don't have to make a speech!

DEPARTMENT OF SPEECH AND DRAMA
UNIVERSITY OF COLORADO
BOULDER, COLORADO

April 12, 1968

Dear

We've been doing some research on you, and the best we can determine, your native language is English. You are a student at the University of Colorado. You are either a graduate or an undergraduate, majoring in engineering or some other subject.

We need your help in a validation study of a new test of English proficiency. So, we are offering you opportunities to

a) make a contribution to the body of scientific knowledge about language,

b) help establish a reasonable standard of acceptable English usage, and

c) receive $5.00 in cash.

All you have to do is take a two hour test. It will be given in Guggenheim 201 on Wednesday, April 17th (7:00 to 9:00 p.m.) and on Saturday, April 20th, (9:00 to 11:00 a.m.). You may choose either one of these times.

If you plan to take the test on Wednesday, just bring this letter with you. If you plan to take the test on Saturday or would prefer some other time, please return the enclosed card telling us of your wishes.

Sincerely yours,

Donald K. Darnell

Donald K. Darnell, Project Director
Extension 7488

p.s. It's a paper and pencil test--you don't have to make a speech!

DEPARTMENT OF SPEECH AND DRAMA
UNIVERSITY OF COLORADO
BOULDER, COLORADO

April 29, 1968

Dear

We are engaged in a validation study of a test of English proficiency. Although the primary object of the study is a better test of proficiency for foreign students, part of the task is to determine how native speakers of English use their own language.

So far we've tested 48 foreign students and 160 native English users. To fulfill our commitment to the U. S. Office of Education, we need 40 more native speakers of English. We need you. And we'll pay you $5.00 for your trouble.

We'll be testing next Saturday morning, May 4th, at 924 Broadway (across from the Country Store). The test takes about 1½ hours (2 hours maximum).

If you can help us out, call 7488 any morning this week or 444-4487 any evening and tell us what time to expect you next Saturday.

Please!

Sincerely yours,

Donald K Darnell

Donald K. Darnell
Project Director

Appendix C

Sample Letters to Foreign Students

UNIVERSITY OF COLORADO
DEPARTMENT OF SPEECH AND DRAMA
BOULDER, COLORADO 80302

March , 1968

Dear

You would probably agree that the program of testing and teaching English as a second language in the United States could be better. We are currently working on this problem with research sponsored by the U. S. Office of Education, and we need your help.

To complete the research project, we need approximately 80 foreign students to take two tests. One test is the Test of English as a Foreign Language (TOEFL), administered by the National Testing Service and required for admission by several major universities. The second test reflects a new concept in testing. It is specifically aimed at determining whether the foreign student has the command of English he needs to pursue his chosen academic program in an English speaking environment.

Our purpose is a comparison of these different testing procedures, rather than evaluation of your individual language proficiency. Your name and scores would not be available to anyone outside the research project except at your request.

This is why we need your help. You have been chosen in a random sample of foreign students on the C. U. campus. It is very important that you participate, because every substitution we are forced to make will challenge the validity of the total study.

The TOEFL will be given on April 13th, a Saturday morning. The costs of this test ($10.00 per student) will be paid by the research project. The second test will be given April 17th, Wednesday evening, or Saturday April 20th, at your convenience. On completion of both tests we will pay you $5.00.

Please return the enclosed postcard immediately. If you indicate that you are willing to help us, we will send you a reminder card telling the place and the exact time of the test administrations.

If you have any questions, you may call university extension 7488 during the day or 444-4487 in the evening.

Sincerely yours,

Donald K Darnell

Dr. Donald K. Darnell, Project Director
Associate Professor
Department of Speech and Drama

Dear

Since we have not yet received your response to our previous letter, we are writing you again to encourage you to participate in the language proficiency testing project.

Although you, personally, may not have difficulties with the English language, some foreign students do. We are trying to help them by improving the procedures for testing language proficiency, but we cannot without your help. If you will help us and the results are as we expect, the results of this study will have a significant effect on testing procedures across the country. Regardless of the result, the information obtained will certainly have an effect on future admission and language requirements at the University of Colorado. But, again, our acquiring any useful information depends to a large extent on you.

If you are benefiting from your education here, this is an opportunity to return the favor by providing information that no one else can give. If you are not, the reason may be that not enough research of this kind has been done in the past.

The TOEFL will be given on Saturday, April 13th in Guggenheim 201 starting at 8:45 a.m. and ending at noon. The second test will be given Wednesday, April 17th from 7:00 to 9:00 p.m. and Saturday, April 20th from 9:00 to 11:00 a.m., also in Guggenheim 201. You may choose either time to take the second test.

If you are willing to participate, please return the postcard with an indication of when we can expect you to take the tests. If you cannot participate, we need to know that too and as soon as possible.

Sincerely yours,

Donald K. Darnell
Project Director

DKD:kl

Dear

You, personally, may not have difficulty with the English language, but some foreign students do. WE NEED YOUR HELP TO HELP THEM. We believe that the program of teaching and testing English is reasonably good, but that it could be better. A significant part of improving that program is the development of better tests of language proficiency which focus on the student's needs.

Under the sponsorship of the U. S. Office of Education, we are currently working on the problem of testing language proficiency. Our project is designed to compare the Test of English as a Foreign Language (TOEFL), which is administered by the National Testing Service, with an alternate testing procedure. If the results of this study are as we expect, they will have a significant influence on language testing across the country. Regardless of the results, the information obtained will certainly have an effect on future admission and language requirements at the University of Colorado.

You have been chosen in a sample representative of foreign students on the C. U. campus. Your responses to two tests would help us decide which one (if either) is an appropriate method for determining who needs help with English in order to get the most from his (or her) educational experience in the United States. We realize that your time is valuable, but we believe that this study is important enough to justify our urging you most strongly to participate. If you are benefiting from your education here, this is an opportunity to return the favor by providing information that no one else can give. If you are not, the reason may be that too little research of this kind has been done in the past.

Although we believe that foreign students are the ones who will benefit most from this research, an exception has been made to federal policy allowing us to give you, on completion of both tests, $5.00 as a token of appreciation for your assistance. The normal test fee of $10 per student is also being paid by the project to obtain this information.

The TOEFL will be given only on Saturday, April 13th in Guggenheim 201, starting at 8:45 a.m. and ending at noon. The second test will be given twice; Wednesday, April 17th from 7:00 to 9:00 p.m. and Saturday, April 20th from 9:00 to 11:00 a.m, also in Guggenheim 201. You may choose either time to take the second test.

Please return the enclosed postcard immediately. If you indicate that you are willing to participate, we will count on you to arrive at the test sessions approximately 5 minutes before the starting times indicated above. Keep this letter as a reminder.

If you need further information, call university extension 7488 or 444-4487.

Sincerely yours,

Dr. Donald K. Darnell
Project Director

Appendix D

Samples of Computer Output from CLOZENTROPY Scoring Program

Note: The responses to blank number 2 of each of the four messages were arbitrarily chosen to illustrate what a hand scoring key would look like. In each case, the responses of the 200 native Ss are followed by the responses of the foreign student sample. The 49th foreign student is, in each case, the author of the text passage. Differences in headings indicate different versions of the program.

The last page in this appendix shows the kind of lists of individual scores that may be obtained. The raw scores from a separate run have been superimposed on the printout of coded scores. These scores are the foreign students scored against themselves on the undergraduate non-engineering message.

GO

```
BLANK NUMBER  2 --- 200 SUBJECTS YIELD 15 DIFFERENT RESPONSES

              ENTROPY SCORING FOR  CLOZENTROPY


       TOTAL INFORMATION                            .8274

            (OMIT)                     0          -6.8164
            APPARENT                   1          -6.8164
            CERTAIN                    2          -5.8164
            CONSTANT                   1          -6.8164
            EVENTUAL                   1          -6.8164
            HIDDEN                     1          -6.8164
            INHERANT                   1          -6.8164
            INHERENT                   4          -4.8164
            MANIFEST                   1          -6.8164
            OBVIOUS                    3          -5.2315
            OF                         1          -6.8164
            POSSIBLE                   1          -6.8164
            SEVERAL                    1          -6.8164
            SOME                       1          -6.8164
            THE                      180            .6754
            VARIOUS                    1          -6.8164


BLANK NUMBER  2 ---  49 SUBJECTS YIELD  6 DIFFERENT RESPONSES

              CLOZE SCORING FOR   CLOZENTROPY


            (OMIT)                     0
            ALL                        1
            HAVING                     1
            MANY                       2
            SIMILAR                    1
            THE                       43
            WHICH                      1
```

UO

```
BLANK NUMBER  2 --- 200 SUBJECTS YIELD  4 DIFFERENT RESPONSES

              ENTROPY SCORING FOR  CLOZENTROPY


         TOTAL INFORMATION                          .1361

              (OMIT)                    0         -7.5077
              ANCIENT                   1         -7.5077
              FOLK                      1         -7.5077
              NATIVE                    1         -7.5077
              OF                      197           .1143
```

```
BLANK NUMBER  2 ---  49 SUBJECTS YIELD  4 DIFFERENT RESPONSES

              CLOZE SCORING FOR   CLOZENTROPY

                (OMIT)                  0
                AND                     1
                LIKE                    1
                OF                     46
                THE                     1
```

```
BLANK NUMBER  2 --- 200 SUBJECTS YIELD  7 DIFFERENT RESPONSES

          CLOZE SCORING FOR  CLOZENTROPY


          TOTAL INFORMATION                       1.9162

                    (OMIT)                0      -5.7276
                    AT                    2      -4.7276
                    BY                   50       -.0838
                    FROM                 25      -1.0838
                    IN                    3      -4.1426
                    INTO                 21      -1.3353
                    THROUGH               1      -5.7276
                    TO                   98        .8871
```

```
BLANK NUMBER  2 ---  49 SUBJECTS YIELD  5 DIFFERENT RESPONSES


               CLOZE SCORING FOR   CLOZENTROPY


                    (OMIT)                0
                    BY                    6
                    FROM                  5
                    INTO                  1
                    TO                   36
                    TO THE                1
```

BLANK NUMBER 2 --- 200 SUBJECTS YIELD 46 DIFFERENT RESPONSES

ENTROPY SCORING FOR CLOZENTROPY

TOTAL INFORMATION		3.9635
(OMIT)	0	-3.6804
ACTION	4	-1.6804
ANYWHERE	1	-3.6804
AREA	1	-3.6804
CHANGE	1	-3.6804
CONDITION	3	-2.0954
COUPLE	1	-3.6804
ELEMENT	2	-2.6804
ELEMENTS	1	-3.6804
ENERGY	5	-1.3584
ENTROPY	1	-3.6804
EQUILIBRIUM	4	-1.6804
FACTOR	2	-2.6804
FORCE	68	2.4071
FORCES	28	1.1270
HEAT	2	-2.6804
MATERIAL	2	-2.6804
MATTER	1	-3.6804
MECHANISM	3	-2.0954
MECHANISMS	2	-2.6804
MEMBER	1	-3.6804
MOTION	4	-1.6804
MOVEMENT	1	-3.6804
NOTICEABLE	1	-3.6804
PART	9	-.5105
PARTS	6	-1.0954
PHASES	1	-3.6804
PLACE	1	-3.6804
PORTION	1	-3.6804
PRESSURE	2	-2.6804
PRESSURES	1	-3.6804
PROCESS	1	-3.6804
QUANTITY	1	-3.6804
RATIO	1	-3.6804
REACTION	2	-2.6804
REGION	1	-3.6804
RELATIONSHIP	2	-2.6804
SECTION	2	-2.6804
SITUATION	1	-3.6804
STATE	16	.3196
STRESS	3	-2.0954
STRUCTURE	1	-3.6804
SUBSTANCE	1	-3.6804
SYSTEM	1	-3.6804
TEMPERATURE	5	-1.3584
TEMPT	1	-3.6804
WEIGHT	1	-3.6804

UE

BLANK NUMBER 2 --- 49 SUBJECTS YIELD 20 DIFFERENT RESPONSES

CLOZE SCORING FOR CLOZENTROPY

(OMIT)	4
AMOUNT	1
DISORDER	1
ELEMENT	1
ELEMENTS	1
ENERGY	1
FACTOR	1
FORCE	13
FORCES	9
IF	1
MASS	1
MOMENT	1
MOTION	1
PART	1
PARTS	1
REACTION	1
STATE	7
STATES	1
STRUCTURE	1
TEMPERATURE	1

LISTING OF INDIVIDUAL CLOZE SCORES FOR CLOZENTROPY

201	.886	-9.139
202	.974	-2.163
203	.632	-30.086
204	.558	-34.613
205	.962	.243
206	1.062	4.100
207	1.177	11.464
208	.915	-6.659
209	1.178	11.510
210	.967	-2.608
211	.855	-11.628
212	.678	-28.296
213	1.061	4.025
214	1.186	11.972
221	1.482	28.027
222	1.682	37.166
223	.921	-6.291
224	.599	-37.340
225	1.457	26.943
226	1.077	5.060
227	1.212	13.566
228	1.508	29.371
229	.459	-56.595
230	.771	-19.020
231	1.033	1.914
232	1.204	12.967
233	.698	-26.370
241	1.207	13.295
242	.670	-29.134
243	.927	-5.709
244	1.322	19.916
245	1.309	19.206
246	1.522	30.071
247	1.399	24.001
248	1.317	19.671
249	.345	-77.090
250	1.014	.673
251	.998	-.361
261	.885	-9.171
262	1.200	12.911
263	1.583	32.930
264	.576	-39.966
265	1.217	13.897
266	1.068	4.330
267	1.125	8.246
268	1.255	16.103
269	1.551	31.427
270	.685	-27.717
301	1.235	14.954

Appendix E

Significantly Different Cell Means--Native Data

Appendix E

Significantly Different Cell Means--Native Analysis

L1 = Graduate level

L2 = Undergraduate level

C1 = Engineering content

C2 = Non-engineering content

M1 = Engineering major

M2 = Non-engineering major

D1 = Easy material

D2 = Difficult material

L1	L2
1.13491	.98998

M1	M2
1.09999	1.02491

	C1	C2
L1	1.17858	1.09125
L2	.96655	1.01342

	C1	C2
M1	1.15444	1.04554
M2	.998379	1.05913

		C1	C2
L1	M1	1.27594	1.06705
	M2	1.08122	1.11545
L2	M1	1.03295	1.02403
	M2	.90015	1.00281

		D1	D2
L1	C1	1.13750	1.21966
	C2	1.10270	1.07980
L2	C1	.97269	.96041
	C2	.99489	1.03195

Appendix F

Significantly Different Cell Means--Foreign Data

Appendix F

Significantly Different Cell Means--Foreign Analysis

L1 = Graduate level
L2 = Undergraduate level
M1 = Engineering major
M2 = Non-engineering major
C1 = Engineering content
C2 = Non-engineering content

D1 = Easy material
D2 = Difficult material
R1 = Graduate criterion group
R2 = Undergrad criterion group
G1 = Engineering criterion group
G2 = Non-engineering criterion

C1	C2
.63473	.58955

R1	R2
.58429	.63999

G1	G2
.56028	.66400

	C1	C2
M1	.69375	.54956
M2	.57571	.62953

	D1	D2
C1	.64442	.62504
C2	.53675	.64234

	R1	R2
M1	.59972	.64359
M2	.56885	.63639

	R1	R2
C1	.58121	.68825
C2	.58736	.59173

	G1	G2
C1	.56524	.70422
C2	.55532	.62377

	G1	G2
D1	.54706	.63412
D2	.57351	.69387

		R1	R2
M1	C1	.65061	.73689
	C2	.54884	.55029
M2	C1	.51181	.63961
	C2	.62589	.63317

		R1	R2
C1	D1	.59637	.69247
	D2	.56605	.68402
C2	D1	.53234	.54116
	D2	.64239	.64230

		G1	G2
C1	D1	.59391	.69494
	D2	.53656	.71351
C2	D1	.50020	.57330
	D2	.61045	.67424

		G1	G2
C1	R1	.50137	.66105
	R2	.62910	.74740
C2	R1	.56246	.61226
	R2	.54819	.63527

		G1	G2
D1	R1	.51294	.61577
	R2	.58117	.65246
D2	R1	.55090	.65754
	R2	.59611	.73021

			R1	R2
M1	C1	D1	.68975	.75995
		D2	.61147	.71382
	C2	D1	.51035	.52457
		D2	.58732	.57600
M2	C1	D1	.50300	.62500
		D2	.52062	.65422
	C2	D1	.55432	.55775
		D2	.69745	.70860

			G1	G2
C1	D1	R1	.52202	.67072
		R2	.66580	.71915
	D2	R1	.48072	.65137
		R2	.59240	.77565
C2	D1	R1	.50385	.56082
		R2	.49655	.58577
	D2	R1	.62107	.66370
		R2	.59982	.68477

ERIC REPORT RESUME

OE 6000 (REV. 9-66)

DEPARTMENT OF HEALTH, EDUCATION AND WELFARE
OFFICE OF EDUCATION

(TOP)

ERIC ACCESSION NO.

ERIC REPORT RESUME

CLEARINGHOUSE ACCESSION NUMBER

RESUME DATE 10-2-68

P.A.

T.A.

IS DOCUMENT COPYRIGHTED? YES ☐ NO ☒

ERIC REPRODUCTION RELEASE? YES ☒ NO ☐

TITLE

THE DEVELOPMENT OF AN ENGLISH LANGUAGE PROFICIENCY TEST OF FOREIGN STUDENTS, USING A CLOZENTROPY PROCEDURE
Final Report

PERSONAL AUTHOR(S)

Darnell, Donald K.

INSTITUTION (SOURCE)

University of Colorado-Boulder, Colo. Dept. of Speech

SOURCE CODE

REPORT/SERIES NO.

OTHER SOURCE

SOURCE CODE

OTHER REPORT NO.

OTHER SOURCE

SOURCE CODE

OTHER REPORT NO.

PUB'L. DATE 10-2-68

CONTRACT GRANT NUMBER OEG 8-8-070010-2000 (057)

PAGINATION, ETC.

67 pages

RETRIEVAL TERMS

English Language Proficiency TOEFL CLOZENTROPY

Cloze Procedure Information Theory

IDENTIFIERS

ABSTRACT

A test combining cloze procedure and an entropy analysis (CLOZENTROPY), which measures the compatibility of a foreign student's English with that of his peers who are native speakers of English, and the Test of English as a Foreign Language (TOEFL) were administered to 48 foreign students.

Comparable reliability coefficients of approximately .86 were obtained for the two tests. Correlation between total scores on the two tests was .838. Analysis of variance confirms that content and difficulty of test material, major of Ss, and level and major of native comparison groups have significant influences on the CLOZENTROPY index of English proficiency. CLOZENTROPY procedure has numerous advantages over conventional types of tests. Its major weakness is its dependency on computer assistance in scoring.